OBJECT CATEGORIZATION FOR AFFORDANCE PREDICTION

A Thesis
Presented to
The Academic Faculty

by

Jie Sun

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
College of Computing

Georgia Institute of Technology
August 2008

OBJECT CATEGORIZATION FOR AFFORDANCE PREDICTION

Approved by:

James M. Rehg,
Advisor and Committee Chair
College of Computing
Georgia Institute of Technology

Aaron Bobick, Co-advisor
College of Computing
Georgia Institute of Technology

Tucker Balch
College of Computing
Georgia Institute of Technology

Henrik I. Christensen
College of Computing
Georgia Institute of Technology

Pietro Perona
Department of Electrical Engineering
California Institute of Technology

Date Approved: 4 June 2008

For mama, papa and Jungeun

ACKNOWLEDGEMENTS

I am fortunate to have Dr. Jim Rehg and Dr. Aaron Bobick as my PhD advisor and co-advisor, who have given me tremendous mentorship and guidance throughout my graduate study at Georgia Tech. I have always been inspired and stimulated in the discussion and interaction with my advisors. Many of the ideas in this thesis can be regarded as from Aaron's persistent pursuit in understanding natural object categorization and Jim's insightful endeavor of framing the problem in affordance learning with the robotics world.

I would also like to thank my additional committee members — Dr. Tucker Balch, Dr. Henrik Christensen and Dr. Pietro Perona, for their helpful comments and critiques at different stages of the work. Tucker and Henrik had also let me use their powerful toys — the two robots which I've been working on. I also appreciate Pietro's insights from the thesis proposal which helped me to focus on the major elements of the work and make faster progress.

I also wish to thank the professors and students here at Tech for the wonderful discussions and collaborations. In particular, I thank Dr. Irfan Essa, Dr. Greg Turk, Dr. Jarek Rossignac, Matt Mullin and also the fellow students and friends Charlie Brubaker, Kai Ni, Huamin Wang, Jianxin Wu, Pei Yin, and Howard Zhou. My graduate student life is made more enjoyable with you.

Finally I thank my parents and my wife Jungeun Shim. My parents have given unreserved support for me to pursue my dream studying abroad and my wife Jungeun has always been beside me with love and support in the past years, good times and difficult times. This thesis work would not have been made possible without their tremendous support, understanding and encouragement. It is to them that this thesis is dedicated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

A fundamental requirement of any autonomous robot system is the ability to predict the *affordances* of its environment. Affordances are latent "action possibilities" that determine how the robot can interact with the objects in the environment, for example the grass affords traversability. Referring to the robot as a sensormotor system emphasizes that it constitutes the *physical capability* to interact with objects which defines affordances, and its *perception capability* to predict the affordances.

A conventional hypothesis is that affordances are perceived directly from appearance without any intermediate representation. In this dissertation, we demonstrate that the direct perception approach can indeed be applied to the task of training robots to predict affordances, with example experiments in nontrivial tasks and environments. This approach however, does not consider that objects can be grouped into categories such that objects of the same category have similar affordances. Although the connection between object categorization and the ability to make predictions of attributes has been extensively studied in cognitive science research, it has not been systematically applied to robotics in learning to predict a number of affordances from recognizing object categories.

In this dissertation, we develop a theory of learning and predicting affordances where a robot explicitly learns the categories of objects present in its environment in a partially supervised manner, and then conducts experiments on that environment to both refine its model of categories and to learn the category-affordance relationships. In comparison to the direct perception approach, we will explore the hypothesis that categories make the affordance learning problem scalable, in that they make more effective use of scarce training data and support efficient incremental learning of new affordance concepts. Another key aspect of our approach is to leverage the ability of a robot to perform experiments on its environment and thus gather information independent of a human trainer. We develop the

theoretical underpinnings of category-based affordance learning and validate our theory on experiments with physically-situated robots.

Finally, we refocus the object categorization problem of computer vision back to the theme of autonomous agents interacting with a physical world consisting of categories of objects, where the "goodness" or utility of categorization can then be well-defined, which is inherently tied to the goal of making property inference in the world. This enables us to reinterpret and extend the Gluck & Corter category utility function for the task of learning categorizations for affordance prediction.

CHAPTER I

INTRODUCTION

1.1 Motivations

A fundamental requirement for a robot to perform any meaningful task is the ability to *interact* with the objects in its environment and the ability to *perceive* the possible actions that it can perform on the objects. The action capabilities enabled by its physical properties and the perception capabilities obtained through its sensors and data processing algorithms constitutes a sensor-motor system [30]. (A detailed discussion with next generation personal robots is in section 1.2.)

This motivates two research topics in computer vision and robotics: object categorization and affordance prediction. The concept of *affordance* was originated by J.J. Gibson [25, 26] to refer to the "action possibilities" latent in the environment. Jointly determined by the agent (or organism) and the environment, an affordance is dependent on the agent's capabilities. Affordances of an object determine how the agent can interact with the entity. These affordances are objectively measurable, independent of the agent's perception capabilities. For example, whether a chair *affords sitting* (for a particular person) defines the *sittable affordance*. Moreover, whether the person perceives that possibility or not does not deny the existence of the affordance.

The task of object categorization or category recognition is to recognize individual object instances as coming from certain categories. In the broad sense, a *categorization* can be any partition of objects, such that objects from the same category are treated equivalently, even if they are distinguishable [54, 30]. Category recognition has always been a major topic in computer vision as an means for image understanding and it has been mainly focused on natural object categories [19, 86].

Although there is great progress in category recognition research, one aspect that has normally been overlooked in the vision community is how categories are defined. This can

be specifically stated as a number of closely related questions concerning (1) the intension and extension of the category, (2) the reason to consider some categories but not others, and (3) the purpose of category recognition other than for the sake of recognizing these categories. The computer vision paradigm is typically a disembodied task in that categories are defined with a set of human labeled images and the goal is to learn the classifier that best *matches the given labeling*. On the other hand, these questions have been extensively discussed in the cognitive science literature, with a body of research arguing that objects are categorized to facilitate prediction of attributes [10, 54, 30]. We adopt this notion of categorization in a affordance prediction setting, where categorization serves only as an intermediate representation of object-partitioning which facilitates affordance learning. When directly connected to the task of affordance perception, the ambiguity of categorization can be reduced: "good" categories are the ones that helps affordance learning and prediction.

The connection between affordance and category is apparent as suggested by the omnipresence of affordances in defining object categories. For example, the definition of cat as "a carnivorous mammal long domesticated as a pet and for catching rats and mice" in Webster's dictionary [1] includes at least 3 affordances: carnivorous, domesticated, and (the ability of) catching mice. Back to the chair example, a person looking for a place to sit, or equivalently the sittable affordance, is more likely to look for the chair category rather than a particular chair instance. As precisely stated by Russell & Norvig in [74], "although interaction with the world takes place at the level of individual objects, much reasoning takes place at the level of categories". This also supports our approach of predicting affordances via first recognizing the category of the object.

In terms of affordance perception, because of the great variability of environments and their appearance, the only plausible way for a robot to acquire such knowledge is through learning via training and experimentation. The question then arises how should an agent learn (or be trained) to predict affordances of objects in its environments from their appearance? Gibson argued that the inference process is immediate and direct — hence the phrase "direct perception" — and as such for each affordance of interest a direct mapping must be learned from appearance to affordance. This massive learning requirement where every new

affordance is a new learning problem was one of the key objections of a variety of rebuttals to Gibson's work [87, 63, 60]. Another concern is that such a learning mechanism does not take into account any of the inherent structure of the world of being made up of particular object types each having their own set of affordances with respect to the observing agent.

1.2 Affordances for Next Generation Personal Robots

In this section, we consider a next generation personal robot that works in the home or office environment. The typical functionality for such a robot would include the following major tasks: (1) *cleaning*, e.g. vacuuming the floor and picking up small objects on the floor; (2) *surveillance*, such as checking whether the door is locked, oven is turned off, strangers in the house, objects displaced or missing; (3) *fetching items*, e.g. bringing a can of soda from the fridge and getting the mails and newspapers; (4) *simple manipulation*, such turning on/off home electrical appliance, moving furniture and locking the doors.

These are considered the basic functionalities for any next generation personal robot, on top of which more advanced ones can be designed. For example, different floor types may require different cleaning methods such as sweeping for the tiles. The description of these functionalities determines the affordances that the robot will be required to learn. First, *traversability* perception is the basic requirement for navigation: the robot needs to predict what part of the home affords traversing. Simply maintaining a 3D map of the house does not work because the environment (e.g. furniture layout) may change. Second, navigation requires moving the *movable* furniture out of the robot's way, this can also gain the robot access to the floor for cleaning. Moreover, object fetching requires the robot to know what objects are *liftable* and more complicate scenarios even require different ways of lifting: heavy objects requires two-arm carrying, fragile objects requires careful handling, a cup of coffee needs to be hold still to avoid splitting – these all define (at finer granularity) different affordances. Furthermore, affordances are defined for doors and windows (*openable* and *lockable*), home appliance (*operative*), etc. While we anticipate the robot to typically consider 10-20 relevant affordances, the number may increase depending on its capabilities and task specification.

The next question in learning is then whether the robot should learn these affordances in a brute-force affordance-by-affordance binary classification fashion. In other words, whenever a new affordance is introduced, the robot spends weeks to collect a large amount observation it may acquire and learn a binary classifier as if learning an arbitrarily pattern but not a real world property. This overwhelming data requirement for learning may significantly prevent home users to appreciate the practical use of the robot. The alternative of a fast learning mechanism from a prior knowledge base of typical object categories is more promising. For example, if the robot has the knowledge that doors are openable and lockable, it may concentrate on learning the new appearance of the doors at a different environment. If there are other object categories that have the same affordance values, they can similarly be learned by updating their appearance model. Then learning openable versus non-openable objects is decomposed into learning the appearance of these object categories. Furthermore, when there are doors that are not openable (e.g. locked), the robot is only required to distinguish the nuance between locked versus not-locked doors in this object category. The difference may differ in different environment and therefore needs to be learned, but the general category-to-affordance properties about the door category and the openable and lockable affordances can be built into the robot's knowledge.

In term of what object categories are necessary to learn, one would typically expect to include the following ones in a home/office environment: table, chair, door, floor, etc. In this work, we won't give a mechanism for generating the list of categories, but will provide a way to compare the goodness of different sets of categories, via the category utility. This makes it possible to learn object categorization through evolution, such that new categorization hypotheses are proposed and compared against prior ones until this trial-and-error process reaches a good categorization for the affordance tasks.

1.3 Contributions

Our thesis is that object categorization can be used as an intermediate representation that makes affordance learning and prediction more tractable. Towards this, in this dissertation we develop a theory of learning and predicting affordances where

a robot explicitly learns the categories of objects present in its environment in a partially supervised manner, and then conducts experiments on that environment to both refine its model of categories and to learn the category-affordance relationships. In comparison to the direct perception approach, we will explore the hypothesis that categories make the affordance learning problem scalable, in that they make more effective use of scarce training data and support efficient incremental learning of new affordance concepts. Another key aspect of our approach is to leverage the ability of a robot to perform experiments on its environment and thus gather information independent of a human trainer. We develop the theoretical underpinnings of category-based affordance learning and validate our theory on experiments with physically-situated robots.

The most important contribution of this work is to *provide a computational model to tie together object categorization and affordance prediction.* To the best of our knowledge, this is the first attempt to systematically frame object categorization as an intermediate representation for multiple affordance learning and to conduct experiments on real world robots.

This is motivated by the observation that typically the world is indeed comprised of categorical entities — be it explicit objects ("bowls, chairs, boxes") or terrain types ("grass, gravel path, puddles") — and focusing learning on these entities is an efficient use of sensor data. Furthermore, because the nature of an object or terrain (i.e. its category) constrains the physical properties, affordances of these categories are not independent or arbitrarily combined. Therefore, information about the category shared among the affordance tasks not only makes small-data affordance learning tractable but also enables more consistent prediction of a number of affordances at the same time. In this research, we refocus the categorization and object recognition problem of computer vision back to the world of autonomous agents interacting with a physical world consisting of categories of objects. The task of category recognition is then inherently tied to the goal of making property inferences in the world. While the models developed here provide particular mechanisms for accomplishing the learning and inference, the broader concept is to re-establish object recognition as a means to an end — namely the ability to infer the important properties

with respect to the capabilities of the robot.

A second contribution is *the use of a robot as an active experimenter as the reference for the discussion of affordances, which is able to test affordances on its own*, thereby greatly leveraging whatever training data has been provided to it by human supervision. We have *demonstrate learning affordance prediction from appearance in a number of nontrivial tasks and environments, rigorously tested in the LAGR project [36].*

In traversability learning for example, the fact that the robot has successfully driven over a patch of terrain suggests that the terrain is traversable. However, instead of learning to recognize the affordance being present after the experiment is taken (such as in [59]), we utilized the affordance observations to learn to predict affordance in the future, i.e. to recognize the action possibility before actually performing it — which is more useful for environment understanding and task planning. This learning to *predict* rather than to *acknowledge* the affordance has long been the underlying theme of robot navigation where the traversability affordance is being studied. We extend this to affordance learning in general by experimenting with a number of affordances.

The third contribution is to *study affordance learning with a categorical representation in comparison to the conventional direct perception approach.* We introduce the Category-Affordance Perception (CAP) model to describe how affordances are defined between the robot and the object of interest as well as the connection between affordance and the object's appearance. This leads to a computational model of a Bayes Net (BNT) which models the probabilistic dependency between the object appearance, the affordances, and the category it belongs to. We explore two types of models, differing in their assumptions about how category recognition affects affordance perception. One model assumes that knowing the category is sufficient to predict affordances. This can be seen as a reasonable direct extension of category recognition in the robot affordance setting, making use of the perception system's capability of category recognition for the task of affordance prediction. In this approach the affordance prediction performance is limited by the category recognition performance and by how well the assumption holds for the environment. The other model breaks this assumption and learns for each category a classifier to predict affordance. Categorization

is used as both a shared representation to connect affordances as well as a means to adopt the divide-and-conquer principle.

We explore learning algorithms for both of these types of models. In our discussion, the derivation and algorithm are presented in a general fashion without assuming particular categorization models, feature representations and learning algorithms, thus making it possible to adopt any state-of-the-art techniques suitable for a given task domain. For the completeness of the algorithm, we explore a variety of issues related to training, such as generative versus discriminative training, direct optimization versus generalized EM, and different ways to make use of the category labels.

The categorical affordance learning approach is shown to outperform direct perception in a number of tasks including: (1) learning with scarce or unbalanced training data, (2) incremental learning of new affordances, (3) making consistent prediction of a combination of affordances, and (4) enabling the prediction of some affordance by measuring other affordances. This justifies the applicability of modeling an object categorization for affordance learning.

Finally, we *provide new insights to the question of whether some categorizations are "better" than others* — in other words, whether there is a "goodness" measure of categorizations that can be used to *rank order* different categorizations. We draw upon previous research of category utility [27, 20] to measure the predictive power of categories in the task setting of affordance prediction. We show that the category utility is directly connected to learning the category-affordance model: its negate serving as an upper-bound of the classification error. Therefore learning to minimize the affordance error effectively maximizes the category utility. We compare the category utility for different categories, on both the training and the testing data as well as the new affordance data, and argue that a more realistic utility measure should take into account the category recognition performance. This implies that categorization optimality is both robot (i.e. sensors and physical capabilities) dependent and task (i.e affordances) dependent. We conclude with a discussion of the implication of category utility on the design of supervised category learning as a means for robot affordance learning.

CHAPTER II

BACKGROUND

2.1 Object Recognition and Categorization

There has been a substantial amount of recent work in computer vision on learning and recognizing object categories. Much of this work has addressed object representations and learning algorithms for reliably discriminating between pre-defined visual object classes. In contrast, the focus of this thesis work is on developing the connection between object categorization and the problem of affordance learning in robotics. As a result of our focus, we do not review in detail the wide range of issues in image feature representation [15, 51], object modeling [66, 7], efficient estimation [85, 62], and classifier design [29, 77]. Instead, we review the previous object recognition literature from the perspective of how object categories are defined.

Typically, an object category refers to a group of object instances that are by definition similar within the group, but distinct to objects outside the group. In the most common approach, object categories are defined implicitly by a manually-assembled collection of images, with each category corresponding to a noun. For example, the widely-used Caltech 101 dataset was created by two subjects who generated object categories by "flipping through the pages of the Webster Collegiate Dictionary, picking a subset of categories that were associated with a drawing" [19], p. 600. The Google Image Search engine was then used to retrieve a subset of images for each category noun. These images were filtered by hand to remove linguistically-correct but perceptually-ambiguous images – "such as a zebra-patterned shirt in the zebra object category". Related manual procedures were used to construct all other popular datasets of general object categories, of which [64, 62, 17] are representative examples. More recently, object category databases have been explicitly constructed to capture variations in pose and lighting. Using these datasets, methods have been developed for 3D object categorization. Recent examples of this approach include [45, 76, 84, 72].

Another approach to the systematic development of object categories is to leverage taxonomic or ontological knowledge. For example, well-defined categorizations can be produced for biological entities such as butterfly wings [48] and leaves [79], by leveraging existing biological taxonomies. This reduces the ambiguity in defining a category but in general such taxonomies will involve attributes that are not visible in camera imagery. An approach to developing general object categories is described in [33], in which the lexical database WordNet is extended with visual descriptions and used to analyze image data. Recent work in [52] extracts semantic graphs also from the Wordnet and trains semantic hierarchic classifiers for object detection. In the classical work on functional object recognition [80], object categories are described by pre-defined functional properties and the recognition is through a process of physical simulation. More recent work in the same spirit is described in [91]. New hardware applications such as RFID-based sensing can also provide new source of information which can be combined with knowledge of human activities to assist learning [98].

An alternative to the use of explicit ontological knowledge is to take a more data-driven approach to the generation of object categories by means of clustering. The work of Barnard et. al. [67, 5] explicitly modeled the connection between images of objects and words. In this approach, images with associated text captions are analyzed to learn a joint model based on latent Dirichlet allocation that links words and image regions. Related work by Russell et. al. has attempted to learn object categories directly from unlabeled image collections via probabilistic Latent Semantic Analysis [73].

As evidenced by previous work, the widespread availability of nontrivial image datasets has resulted in substantial progress in the representation and recognition of object categories. Our goal in this thesis work is to leverage this progress and explore a systematic and constructive approach to the definition of object categories in a task driven setting with affordance learning for a referencing robot. Linking category recognition back to robot perception makes it natural to think of category recognition no longer as an end task by itself, but rather a means towards affordance prediction — more directly connected to the robot's perception of the environment and its planning of task execution.

2.2 Object Categorization Utility

Finally, we note that it has been proposed in the cognitive science literature that the purpose of categorization (i.e. object category recognition) is for prediction, inference and decision making, which are all closely related with the affordance concept. The categorization utility work by Gluck and Corter [27, 12] and by Fisher [20] evaluate the goodness of a categorization by the predictive power of the categories on the feature values, most of which can indeed be regarded as affordances. Bobick argues that the goal of categorization is to be able to predict unobserved properties from observed properties [10]. Harnad argues that "to cognize is to categorize" [30]. A recent review of category utility can be found in Witten and Frank [94]. Although most of the these efforts focuse on the understanding and searching of basic categories, they all emphasize the connection of categorization to prediction of certain properties, where recognizing the object category is not the *ultimate* goal.

In terms of a robot agent performing certain tasks which requires predicting properties of the environment, we suggest that most if not all the properties that it needs to predict are in fact affordances. It would make very little sense to predict properties that are easily measurable, say the height of the object, when stereo equipment is indeed available. In this work, we use object categorization as an intermediate representation for affordance prediction, and also use different affordance values to guide the process of categorization.

2.3 Learning in Category Recognition

An indispensable element in the data-driven category learning framework is the use of machine learning techniques. The models can be roughly divided into two classes: a generative model or a discriminative model. The distinction depends on whether the approach models the extension of the category explicitly (generative) or forms a mapping from observation to the category labels directly through what is called a classification function (discriminative). Gaussian mixture model is a typical example of generative models while a decision tree classifier is discriminative.

A related element in recent investigations into object recognition is the use of combinations of generative and discriminative *training* with a (usually) generative *model* (see for example [32, 46, 61, 89, 35]). From a probabilistic modeling perspective, generative training aims at maximizing the joint likelihood of observation and the class labels, while discriminative training aims at maximizing the conditional likelihood of class labels, which is more directly connected to the classification performance. An alternative discriminative training method is to perform boosting on (generatively) trained models from iteratively re-weighted data distributions [38].

In our categorization-based affordance recognition, we adopt generative model for the category-appearance model so that it can be shared among different affordances. Adopting a generative category-appearance model also makes it possible to incrementally add new object categories with the pre-learned category models unaffected. Generative models also enable augmenting labeled training data with a much larger set of unlabeled data [46]. On the other hand, affordance classifiers from appearance are learned within each object category, i.e. the classifiers can be regarded as "indexed" by the discrete category labels. Training of these two models can be decoupled with an EM framework as we discuss later in chapter 5.

Another relevant research topic to our work is the use of latent variables in graphical models [37, 40], that are introduced to facilitate the representation of the joint probability density over a set of variables. Latent variables are treated as unobserved quantities which makes it easier to specify and manipulate the joint density [8, 31]. Examples of its applications to the object recognition domain can be found in [73, 18, 44]. In our categorization based affordance perception model, the categories can be considered as latent variables which are used to model the joint density of affordance and appearance. We argue that if objects categories are indeed present in the environment, then taking advantage of this categorical structure can greatly improve our ability to learn with limited amounts of training data. Intuitively speaking, appearance space is partitioned into categories, and affordance learning is carried out on a per category basis in a divide-and-conquer fashion. It is important to note that this depends on the existence of object categories and does

not imply that a general procedure of data organization and divide-and-conquer that can improve *any* learning task.

2.4 Affordance Learning in Robotics

The concept of affordance was originated by J.J. Gibson [26] where he defines affordance as jointly determined by organism and environment. Although Gibson argues for a direct perception process for affordance, others have proposed objections and alternatives [87, 60].

While there has been previous work on affordance learning within the robotics community, it has primarily been focused on traversability learning for robot navigation [4, 6, 34, 65, 70, 78, 90], because obstacle detection is essentially the detection of "traversability" or the traversable affordance. An exception is the work on affordance learning in robot manipulation described in [21, 81], which does not make use of visual categorizations. In [24] the liftable affordance is learned with a decision tree on the image patch size and the shape of the object detected by SIFT features. A series of work in [50, 58, 59] tackles a different task of recognizing the affordance *after* the experiment *had* been performed such that the robot action, the effects (both on the object and the robot) and other necessary features are all observed, on top of which a BNT is learned. However, this does not address the problem of predicting the affordance *before* making an action, using sensor data alone.

Early work on learning for navigation was based on manually-acquired training data. Examples include road-following [68, 69] and the detection of open areas using learned visual features [39, 55]. While these efforts deal effectively with some class of terrain features, they are insufficient to address the broad range of terrain and vegetation types found in natural outdoor environments. For example, it would be difficult to use these methods to learn that tall grass or small bushes are traversable. This motivates the design of on-line learning of affordances without human supervision. [88] describes an on-line learning method for indoor robot navigation based on color modeling and stereo sensing. A similar online navigation work in which robot simultaneously collects training data and learns the "true height" of terrain model through real-time interaction with the outdoor environment appears in [93].

Some recent work [75, 13, 14] from the same group has been formalizing affordance as

a 3-tuple of "entity, behavior and effect" for learning robot control. However, the study concentrates on most elementary affordances: the ability to move forward with different angles is defined as a set of completely different affordances. The use of affordance learning as a means for robot control is also the theme of the EU-MACS project, a summary of a related seminar can be found in [71].

The related work in robot control and navigation suggests the connection of the ecological affordance theory to the modern robotics research and its importance. In contrast to previous work on robot navigation (i.e. traversability affordance), we propose to develop a general methodology for affordance learning through a categorization structure of the objects the robot interacts with. We also systematically study the improvement over the conventional direct perception approach in a number of tasks including multiple affordance learning and learning a new affordance with limited training data.

CHAPTER III

AFFORDANCE LEARNING WITH DIRECT PERCEPTION FOR ROBOT NAVIGATION

In this chapter, we present our work on two tasks on ground navigation for an unmanned vehicle (UMV), both of which are framed as affordance learning tasks. The first task is a conventional obstacle detection problem for robot navigation. Instead of assuming the availability of sufficient prior knowledge to define the heuristics as to what constitutes an obstacle, we propose a method by which the robot autonomously collects training data in a new environment and learns the obstacle model online. This provides sufficient flexibility for the robot to learn to adapt to an unknown environment. We demonstrate a trial-and-error approach in which the robot learns from every instance of obstacles that it has encountered as well as every terrain patch that it has successfully driven over.

In the second task, the robot learns from human operated driving examples about which terrain is preferred by a human operator. This is the major theme of the DARPA LAGR (Learning Applied to Ground Robotics) project in the "Learning from Examples" tests [36]. Specifically, a teacher manually drives the robot on a mulch road, emphasizing that the road is more preferable than other open ground, although both of them are equally traversable. This "roadness" or "preferability" is indeed an affordance which it is difficult for a robot to collect training data exploring on its own. Therefore, supervised learning can be helpful. In our approach, rather than assuming that the form of preferred terrain is known *a priori* (e.g. road) and build a specific (road) detector, we explicitly label terrain as preferred versus not preferred. Then a binary classifier is learned to predict preferability. The robustness of the approach has demonstrated in several testing scenarios including that of the LAGR test.

These two specific tasks in robotics demonstrate what is an affordance and how it can be learned. Further more, in both of these tasks, we adopt a direct perception approach. While the affordances we considered are both related to the driving capability of a robot,

we demonstrate the success of our affordance learning framework in the challenging setting of an outdoor, unstructured environment. All of these provide an excellent basis for further discussion with a richer but more structured and controllable context of affordance learning.

3.1 Learn to Predict Affordance: An Example

Consider the scenario of a robot at the entrance of a tunnel, trying to predict whether it can successfully go through. If the robot has knowledge of its own 3D dimensions and is equipped with accurate laser sensors, then recovering the 3D geometry of the tunnel is sufficient to infer whether "in principle" the robot can go through the tunnel or not. In this case, the robot perceives the affordance because it has *sufficient knowledge about the physical properties that defines the affordance, either obtained a priori or through direct measurement with accurate sensors.* If the robot does not know its own size, but does have the laser sensors, then the "allowing passage" affordance is still *a deterministic function* of the 3D geometry of the tunnel, but the parameters are to be learned. Through sufficiently learning, the robot can learn *a mapping* of the 3D geometry to the allowing passage affordance. If however, the robot only has noisy stereo instead of the laser sensors, the robot can similarly learn (but with more effort) a mapping from stereo returns to the affordance, assuming the stereo output is *correlated* with the 3D geometry and hence the affordance. In the case where the robot has neither laser nor stereo but only color camera, it may still attempt to learn the affordance prediction, albeit with probably much greater difficulty and lower accuracy.

Defined as an action possibility, training data of an affordance can be collected through experiments of conducting the action. An experiment in the general sense, is also a type of sensor. In contrast to other perceptual features, the amount of information obtained from an experiment is one bit – true or false. Experiments can be regarded as direct measures of affordances, which is still possible for noise. The fact that the tunnel has allowing passage affordance only in principle suggests that the robot can go through the tunnel, because of the difference between "affordance" and "experiment result". What the robot really predicts is the "outcome of an experiment" as whether it can drive through or not in a

driving attempt, but the actual outcome of a *particular* attempt might be affected by other factors. For example, the robot may drive to the tunnel at a wrong angle or the wheels may get stuck with some trap (such as grass), both will result in a failure of the experiment. It should be emphasized that the affordance still exists: had the robot drove in the right direction and encountered no obstacles, it can successfully drive through the tunnel. In practice we assume these exceptions are rare and therefore can be neglected with more training data being collected.

3.2 Affordance Learning and Direct Perception

Obstacle detection, or traversability prediction is an important component of ground robot navigation. In the common paradigm, an occupancy grid [16] is built as an representation of the world, such that each "cell" in the grid corresponding to a "patch of ground" can be labeled as open space or with obstacles.

Most previous work on outdoor navigation either characterizes the concept of traversability *a priori*, or learns the concept of traversability from manually acquired data [6, 34, 65, 70, 78, 90]. The classification functions are hand-designed based on knowledge of the properties of terrain features such as 3D shape, roughness, image edges or discontinuities.

Although more informative sensors can be made available for the robot, in practice it is still difficult to specify what type of terrain is traversable. First, traversability is a complicated function of *physical properties* of the object, including size, thickness, stiffness and other properties that are difficult to measure or even to define. It is also robot specific, depending on its size, wheel type, driving power and other physical properties that are difficult to distinguish. In terms of sensing capabilities, even a very accurate 3D laser range sensor is not able to obtain sufficient information to assert the traversability of a patch of terrain. For example, the 3D geometry of tall grass can be mistaken for an obstacle because it "occupies" the 3D space that the robot needs to traverse, but the grass may indeed be traversable. On the other hand, tall bushes with similar laser profile can be non-traversable. Therefore, the use of geometry alone to determine traversability is at best a conservative heuristic, avoiding terrains which may be perfectly traversable. Moreover,

because traversability is robot specific, we note that the driving attempt into the same grass by different robots might result in different outcomes. Because of this, it is even difficult for a human to tell beforehand whether the robot is able to traverse a particular patch of terrain. Furthermore, the difficulty is compounded by the wide variety of terrain types such as trees, rocks, tall grass, logs, and bushes. As a result, methods which provide traversability estimates based on predefined terrain properties such as height or shape will be unlikely to work reliably in unknown outdoor environments.

In our work [83, 43], we define obstacle detection as an affordance prediction problem. We develop a framework that allows the robot to collect traversability training data autonomously in an unknown environment through exploration. In other words, the robot experiments by itself the traversability of terrain and learns from the affordance labels it collects. This approach eliminates much of the ambiguity in the definition of an obstacle. We learn a direct functional mapping from appearance to affordance online without any intermediate representation, based on the *direct perception* approach proposed by Gibson.

Although certain physical properties are directly connected to the formation of the traversability affordance, in general they are not directly measurable. Therefore, affordance prediction is an inference process that assumes the appearance (including imagery and all other sensor measurements) is "informative" – although not "deterministic" – of the affordance. In other words, a basic necessary assumption affordance to be predicted from these measurements is that the mutual information of measurements and affordance should be sufficiently high.

We believe that there are three major issues to consider in designing an affordance learning procedure with the direct perception (DP) approach:

1. *Feature Sufficiency:* Since the DP approach attempts to learn a direct mapping from appearance to affordance, the appearance should be a sufficient information set for the affordance to be learnable. Rather than stating it as a *learnability assumption* of the affordance, this is in fact a requirement of the feature space of the appearance. The feature space is considered as "sufficient" if the affordance can theoretically be learned to achieve a pre-specified degree of accuracy. Mathematically speaking, the

Bayes error of an affordance A predication based on appearance X should be less than a predefined acceptable error level:

$$\mathbf{E}\,[\ |A - \mathbf{E}[A|X]|\ ] < \eta \tag{1}$$

2. *Source of affordance labels:* In training, some affordance labels have to be given, either from manual labeling or automatic discovery by the robot. Typically, labels for both the positive and negative affordance values should be provided. But a special case in the second task of "preferability" learning, we demonstrate an example of data labeling through some heuristics when the negative (non-preferable) labels are not provided explicitly.

3. *Data association:* The third important component is to associate the collected affordance labels to the appearance. While an obtained affordance label refers to a particular object or patch of ground terrain, we want to associate this label to the observed appearance of the object. Error in data association can contaminate the training data, labeling what is in truth positive to be negative and vice versa. This is further complicated by sensor noise and localization error.

In the following discussion of the two tasks in traversability learning and preferability learning, we will explicitly address these three key aspects in learning.

3.3 Traversability: Obstacle Detection as Affordance Prediction

Our goal in traversability learning is to develop an automatic training paradigm that required no human labeling of sampled images. We adopt an *on-line learning approach* in which the vehicle learns the affordance of traversability through guided interactions with terrain features and by establishing a correspondence between the vehicle's navigation experience and the visual terrain data acquired from stereo sensors. Navigation experiences such as successful traverses, slippages, and collisions are assessed automatically by means of on-board sensors such as Inertial Measurement Unit (IMU), motor current, and bumper switch. Successes and failures of the navigation provide positive and negative traversability labels for cells in a grid-based representation of the terrain surrounding the vehicle.

By establishing the correspondence between labeled terrain cells and the visual features in the stereo imagery, the robot can automatically label visual features that correspond to traversable and non-traversable terrain examples. The automatically labeled data provide the input to an on-line classifier learning algorithm. As the robot explores its environment, the classifier is trained incrementally. At any point in time, the classifier can make predictions about the traversability of the visible terrain based on its past experiences. These predictions are used to continuously plan an optimal path to the goal. Compared with a predefined heuristic approach, our approach is more broadly applicable to any environment in which the robot can be safely driven as it uses the interaction between the vehicle and its environment to ground the problem of traversability classification.

3.3.1 Key Assumptions for Traversability Learning

Our assumptions to address the three issues of affordance learning with direct perception in section 3.2 is presented as follows. First, we assume that visual features derived from stereo vision and color imagery are sufficient to discriminate between terrain regions from the standpoint of the traversability affordance. Specifically, we assume that two terrain locations with similar visual features have similar traversability. Note that this does not imply that all traversable terrain locations look similar – there is no simple description of traversable regions in terms of image appearance or geometry. As is standard in classification problems, the question of whether two regions with different traversability properties can be discriminated will depend upon the representational capabilities of the sensors and the features. While it is always possible to identify extremely challenging situations, like logs or stumps hidden behind dense grass, we demonstrate that features based on image texture and stereo ranging are able to discriminate between the several terrain classes commonly encountered in outdoor navigation.

Our second assumption is that we can obtain traversability information about terrain regions without endangering the success of the robot's mission. We employ a standard suite of navigation sensors, such as IMU and motor current, to assess whether the forward motion is successful. We build a occupancy grid [16] of the environment and label the grid

cells over which the robot attempts to drive as either traversable or nontraversable based on the navigation sensor values. Note that this traversability labelling depends critically upon the particular characteristics of the vehicle – a large vehicle may be able to drive over small saplings that would present an insurmountable obstacle to a smaller vehicle. This is exactly why traversability should be treated as an affordance and not simply as a predefined property of different types of terrain. This exploration process also requires that the test of traversability will not be fatal, such as driving into a lake or becoming trapped on a bush and unable to back-off.

The third assumption is that we can establish an accurate correspondence between terrain regions on the global physical world and visual features in the image using stereo information. We use this correspondence to associate traversability labels with the visual features of the corresponding terrain structures. For example, as the vehicle drives towards a patch of bushes, it will image the bushes from several different viewpoints, resulting in a collection of visual feature measurements that correspond to a single terrain object. When the robot finally encounters the bush and determines that it is traversable, this label can be propagated to all of the corresponding feature measurements already collected. Therefore, errors in localizing the robot to its local terrain environment will impact the ability to establish this correspondence. Although it is always good to have accurate localization and pose estimation (such as using SLAM), we demonstrate that a standard GPS-based localization scheme is sufficient for many outdoor situations.

3.3.2 Online Learning with Autonomous Data Collection

There are two key elements of the traversability learning algorithm: a mechanism to autonomously collect and label training data as the robot explores its environment, and an online learning algorithm to update the concepts for both traversable and non-traversable terrains. The data collection method can intuitively be illustrated in figure 1. The robot keeps whatever image patches that it sees in a data pool, with each image patch being an observation of the terrain that occupies a ground cell. Initially all of these data are unlabelled, because the robot hasn't interacted with the terrain yet, so that its traversability is

(a) (b)

Figure 1: Traversability data collection online (a) Data collection at time t_0 where observation vectors for ground cells within a history of duration T are obtained and stored (b) As the robot acquires traversability information about the ground cells at $t_0 + dt$, those associated observation vectors that were previously collected are now labeled as traversable or non-traversable.

unknown. Later in time, the robot will try to go through some of the ground cells that it previously observed, thus uncovering the underlying traversability of the occupying terrain: those that the robot actually drive over are traversable — positive examples, while those that hinder the robot motion are non-traversable — negative examples. To summarize, data are collected as the robot moves around and *labels are obtained at a time lag.* In order to reflect the newly acquired training data in the learnt concepts, we would naturally require an on-line learning method with the advantage that the concepts are incrementally updated with the new data only, rather than to be relearned in an offline fashion with all the data. The advantages of this on-line approach as opposed to an off-line approach is to reduce both the computation cost as well as the memory space that needs to be maintained to store training data. In practice, only the data collected within a fixed time window are stored, both to save the storage space and to alleviate the registration error that can result in mis-correspondence, which can result from GPS error or wheel slippage and typically accumulates through time.

Online learning is based on nearest neighbor and vector quantization. We do an online clustering of the training data for both traversable and non-traversable ones separately. A new training sample is either considered to be in the existing clusters or to form a new cluster if its distance to existing clusters are above a threshold. During classification, a ground cell is classified through a majority vote of the individual classification results of all

the associated observations, as the ground is normally view multiple times. Each individual observation is in turn classified by matching with the closest clusters from both traversable and non-traversable and compared the relative frequency of observations.

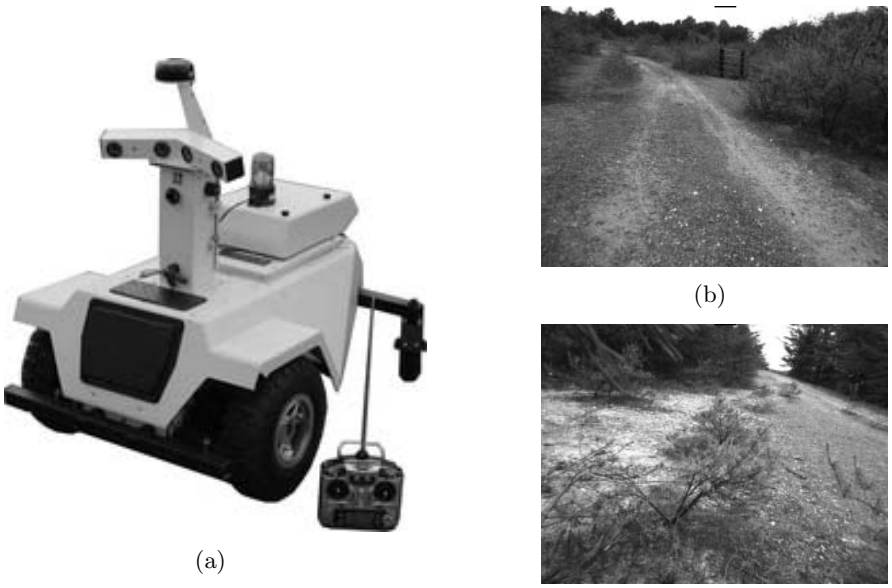
In order to present the implementation details and the experiments, we first provide a brief overview of the robot system that we use.

3.3.3 Robot Architecture Overview

Experiments for both the traversability and the preferability learning are carried out on the LAGR robot in an outdoor environment. Since affordances depend on the robot's capabilities and the appearances measured the sensors, it is worthwhile to briefly describe the robot platform. Depicted in figure 2, the LAGR robot is equipped with four color cameras, both front/rear bump switches, a Garmin GPS receiver, and an IMU. The cameras are paired together so that each pair provides stereo depth maps with a range of up to 10-12 meters (although its accuracy degenerates beyond 6 meters range). The robot is 1.0 meters long, 0.6 meters wide, and 1.0 meters high, with a weight of approximately 60 kg. It can be controlled from a radio-controller or running in its autonomous mode with the behaviors implemented on its imbedded PC. In terms of driving capability, the robot is generally powerful enough to drive on tough terrains and through some dense grass or bushes [82].

The robot carries four computers (2.0 Ghz, Pentium-M, 1 GB RAM) running linux. The central computer, called the *planner computer*, runs mapping, control and planning processes and it can be accessed from outside computers via wireless. The other three computers are connected to the planner computer via ethernet. Two of them are the *eye computers*, each control a pair of stereo cameras for image and stereo processing. The fourth computer carries low-level function including radio-control interface and GPS/IMU integration. Additional detail of our system setup for the LAGR robot can be found in [96].

All of the perception processes are carried out on the eye computers. On each eye computer, stereo depth maps are calculated from the color images collected by the cameras. Given the intrinsic and extrinsic parameters of the cameras and the robot's global position, we can transform a local terrain map onto a global coordinate frame (figure 3). This



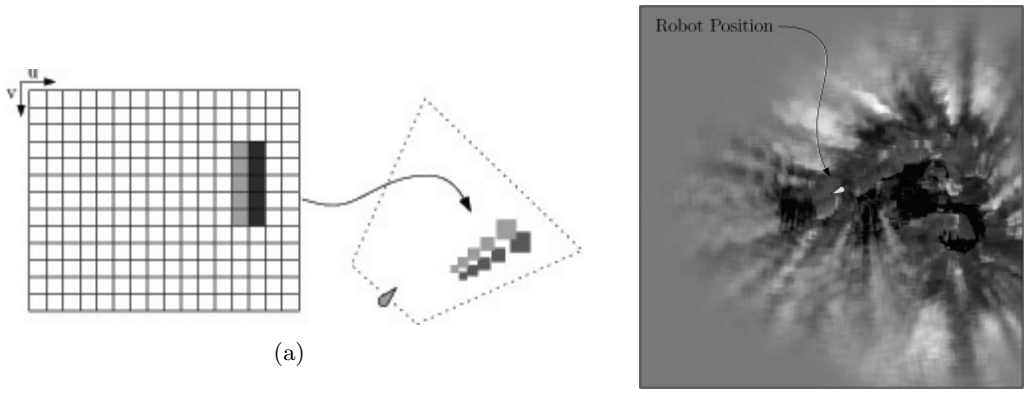
(a) (b) (c)

Figure 2: (a) The LAGR Robot (b)(c) Example rectified images from the stereo cameras. Images show that the robot is capable of driving in a tough outdoor environment and it is powerful enough to drive through bushes. The variety of terrains that the robot can or cannot drive on is hard to define *a priori*, which makes the traversability learning task difficult.

information is subsequently communicated over the network to processes on the planning computer. On the same eye computer, the color camera images are also used to predict the traversability of a terrain patch of the physical world on the global map based on the corresponding image patches observed. The stereo and traversability processes run steadily at 4Hz.

The planner computer receives the two streams of stereo and preferability information from both eye computers and incorporates it into global maps of terrain and preferability, similar to that of an occupancy grid [16]. Each cell on the grid represents a square patch of ground with a length of 0.1m. Under the assumption that the newest information is the more likely to be correct (at least in the vicinity of the robot), past cell information is overwritten with the new. A separate map stores the locations where the robot has encountered hits on its bumper, spikes in its motor amperes, and detections of wheel slippage. These essentially provides the training signals of traversability.

The global maps are the robot's representation of the environment on top of which the



(a) (b)

Figure 3: (a) Traversability information of each image patch is projected on to robot's local map based on the camera intrinsic and extrinsic parameters. (b) The local map is then integrated in to a global traversability map given the pose estimation of the robot.

path planning is performed. In particular, we adopt a heuristically improved A^* planner, or the combinatorial planner described in [95].

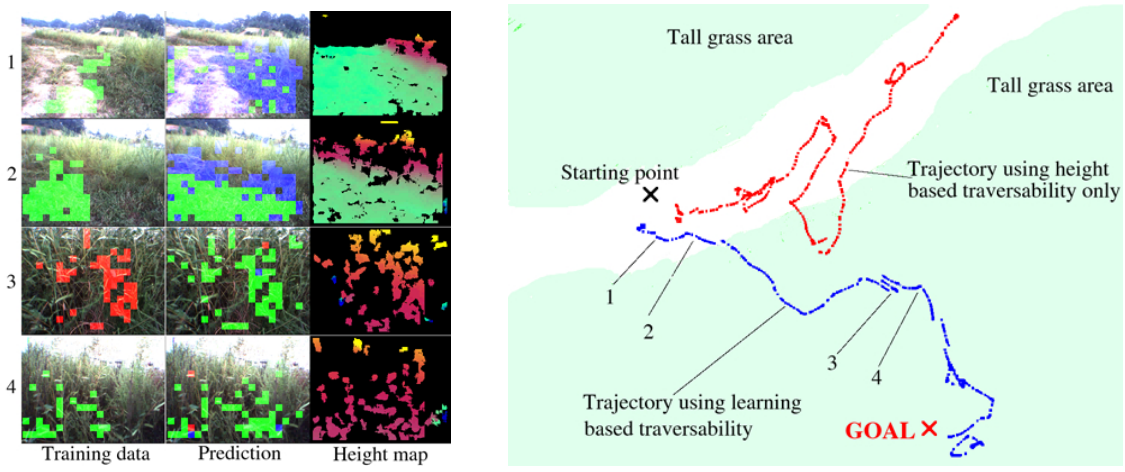
3.3.4 Experiment Results

For this experiment, we only use gray scale images. We do a patch based traversability learning where the image is divided into patches with a fixed size of 12×12 pixels. Computed from stereo depth map, the height estimates of each pixel in an image patch are used to build a 5 bin histogram, forming a height distribution of the corresponding ground patch. Another 8 dimensions of the feature vector are built from a texture histogram based on the Maximum Law's Mask [47]. We note that in our experiments stereo and color information provides the best performance, and was implemented in subsequent LAGR experiments including the Learning from Example test (section 3.4).

Online clustering is based on the χ^2 distance metric of two distributions and the distances of two parts of the feature vectors are weighted by different constants. Within each one feature component, the distance for two feature vectors x_1 and x_2 can be calculated from (where d denotes each dimension):

$$D(x_1, x_2) = \sum_d \frac{(x_1^d - x_2^d)^2}{x_1^d + x_2^d} \tag{2}$$

We focus our discussion on an experiment where the robot is required to drive into tall



(a) (b)

Figure 4: (a) Robot trajectory during two experiments. Online affordance learning (blue path) correctly learns that tall grass is traversable and reaches the goal while a standard height-based method (red path) fails. (b) Rows showing robot view corresponding to the 4 locations on the blue path. Columns showing training and prediction labels and the stereo height map of the image (see text for discussion).

grass in order to reach the goal, more results and discussion can be found in [43]. Two experimental runs are performed: one based on a conventional, pre-specified, height-based traversability classifier where objects above a certain height (as perceived by stereo) are considered obstacles; the other based on our online traversability learning approach. The illustrative map of the environment, the starting and goal points, and the robot trajectories of the two experiments are presented in figure 4.

Figure 4 clearly shows that our approach learns that tall grass is indeed traversable *despite* the fact that its height distribution indicates a possible obstacle. With the traversability model continuously updated online, the planner successfully suppressed the given elevations obtained by the stereo and therefore the robot could traverse the tall grass area to reach the goal. In contrast, the conventional approach went along the edge of the grass area, kept searching for alternative paths where the stereo does not signal alarm. Note that the path of online learning approach (blue) does not lead straight to the goal, since there were some particularly dense tall grass on the way to the goal which proved to be non-traversable for our vehicle. The process of encountering and avoiding these regions resulted in a curved trajectory terminating at the goal position.

The labels of number 1-4 in figure 4(a) indicate the position of the robot where each image in figure 4(b) was captured. The images in rows 1 and 2, captured at the beginning of the run, demonstrate that the online learning method learns that grass is traversable as it had successfully driven through some of it. Blue patches in the middle column of images refer to the case we discussed earlier where no reliable traversability predictions can be made because the observation vector is far from the learned clusters. The images in row 3 are captured at the moment before the robot hits a tall grass area which is too dense for the robot to pass through. The visual appearances of the non-traversable patches are similar to the previous trained traversable patches. The inability to learn the two conflicting grass traversabilities can be attributed to a lack of representational power in our choice of feature space. Moreover, in this case, it is even difficult for human to distinguish between traversable area and non-traversable area. However, we expect in general that similar terrain appearances imply similar traversability properties, allowing our method to be widely useful in practice.

3.4 Preferability: Affordance Learning from Examples

In the second task, we consider a situation in which a human operator is driving a robot through an unknown outdoor environment. One can typically expect the robot to stay on easily traversable regions such as road or other flat ground, while staying away from dense vegetation and mud. A natural objective is to have the robot mimic the human-operated behaviors and maneuvers in a completely autonomous manner. However, the problem is not that of reproducing the behavior at the level of trajectories (e.g. by simply storing the path that the human-operated robot took) but rather to learn a model of the *preferability* of different terrains. By classifying terrains as preferable versus non-preferable, the robot can then plan to drive on preferable terrains whenever possible, therefore acting in a manner similar to the human operated examples.

For example, the human operator may prefer to drive the robot on a road rather than on other type of terrain. The distinction comes from the fact that roads are normally designed to facilitate driving to a specific goal, while other terrain might lead the robot to some areas of difficult driving conditions, such as an area with densely occupied obstacles. This difference clearly implies a special affordance of the road, albeit difficult to collect the affordance information by the robot alone, which justifies the necessity of a human teacher. We refer to this affordance "preferability".

The outlined "Learning from Example" problem is motivated by the DARPA LAGR project [36], in which this problem is an explicit objective of the test. We present our approach for the task, which is based on learning the preferability affordance. Unlike approaches that attempts to build a specific road detector, our approach, based on less constrained assumptions about human preference, is widely applicable to learn other human preferred terrain types with the same fashion.

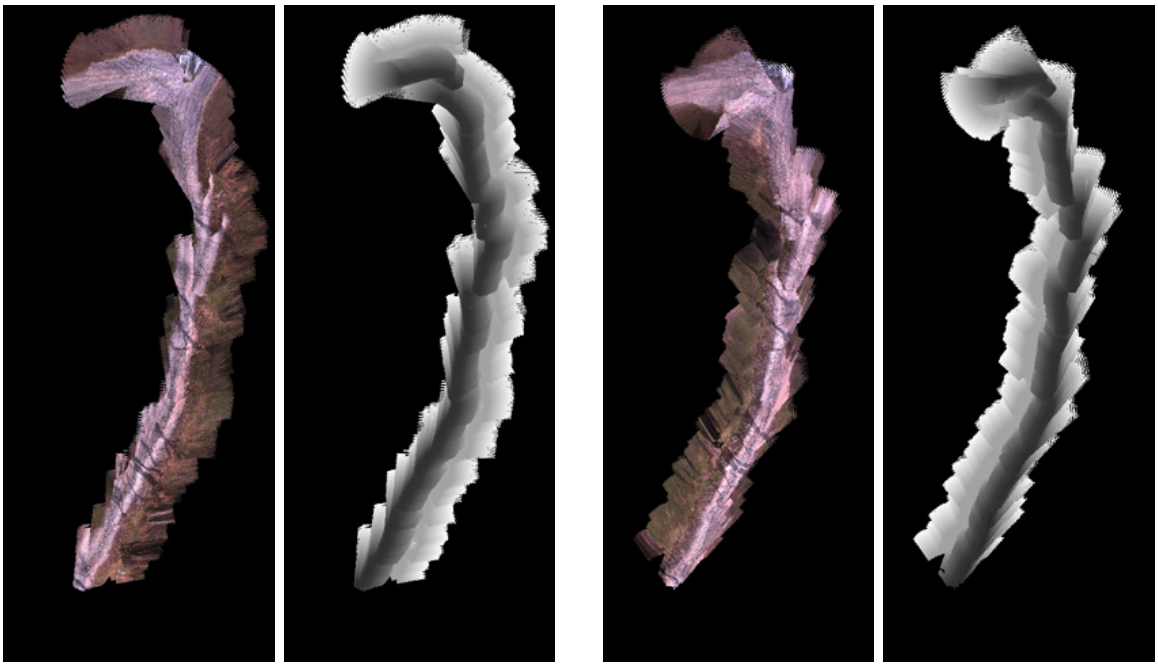
3.4.1 Supervised Learning of Preferability

The experiment setting for preferability is presented in the context of LAGR test 12 "Learning from Example". The same robot system described in section 3.3.3 is used. We focus our discussion here on the visually learning of preferability, while the overall architecture

of the system as well as the integration with controlling and planning can be found in our journal article [82].

As we discussed above, the difference of the preferability learning, as compared to the traversability learning task we discussed before, is that besides avoiding non-traversable terrain regions, we also enable the robot to select preferred regions among *equally traversable* terrain candidates, based on the pattern learned from human operated examples. Although we've described a system that learns terrain traversability on the fly when the robot moves in the environment, there is notable differences between preferability and traversability. While non-traversable terrains are certainly not preferable, traversable terrains are not guaranteed to be preferable. As an example, consider a robot driving to a goal point that is behind a large area of bushes. Assuming there exists a road around the bushes, the robot should ideally take that road to circumvent the bushes, since the road is expected to be safer routes of travel than other open spaces which may lead to difficult driving areas with ditches and obstacles. The direct way planning through the bushes, although shorter, is less preferable. A more intelligent system should prefer the longer path in order to achieve both efficiency and consistency. (Generally speaking, a robot that travels in less time is more efficient; the variance of the travelling time across runs determines consistency: large variance implies inconsistent performance and therefore considered as high risk.)

In this regard, it is natural to have a teacher that instructs the robot the concept of "preferable" terrains by way of providing training examples. Specifically, the teacher manually drives the robot in some similar environment, and the image sequences and robot state obtained by the cameras are stored in a log file. The learning procedure consists of two parts: data labelling and model learning. We use a *ground projected feature map* to obtain the training data. For clarity, we explain our approach with a simplified color-based classifier. The benefit of a color-based classifier is that color of the terrain (as opposed to texture) does not change when observed at different distances. Note that this discussion is meant to be general: both the feature space and the classifier algorithm can be replaced for a different application.



(a) (b) (c) (d)

Figure 5: Ground projected feature maps and the corresponding distance maps are shown for both the left eye (a)(b) stereo and the right eye (c)(d). Dark regions in the distance map correspond to a smaller distance (except for the darkest background region denoting lack of information). The brightest regions are observed at a distance of 10 meters. These 2 pairs of maps are then combined together to obtained final feature map for training.

3.4.2 Ground Projected Feature Map

The main idea of the feature map is to project features of image patches to the terrain from which they come, and store them in the world frame similar to the occupancy grid. For example, for the pixel-wise RGB color feature, we project each pixel from the camera image to the global ground plane, as shown in figure 5(a). Such projection is calculated through the stereo depth map, by first projecting to the 3D location relative to the robot based on the extrinsic and intrinsics of the cameras, and then to the coordinates in the global map based on the robot localization and pose estimation. We maintain a distance map in association with the feature map, such that the distance to the robot at which the feature had been observed is recorded (figure 5(b)).

For our particular application, as new images are obtained when the robot moves, the feature map is over-written when the ground cell is observed at a *closer* distance. At

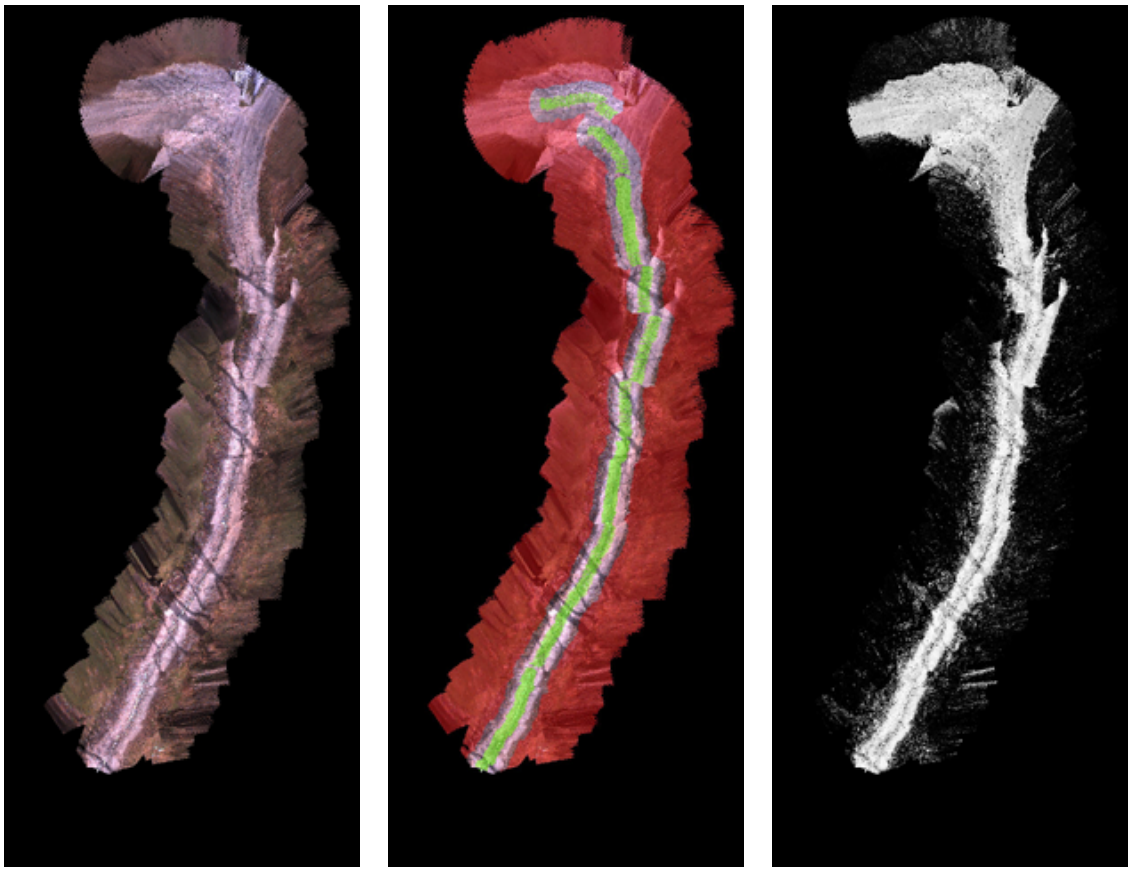
the same time, both the feature map and the distance map is updated, reflecting the new observation. In this sense, such a feature map is called a *minimum distance feature map*. It is also possible adopt this method for a patch based learning with other feature representations (such as texture). In that case, it can be beneficial to construct the feature map with each cell being associated with observations from different distances. Compared with the minimum distance feature map, the full-distance-profile of the terrain can be informative for classification. Other variations such as feature projection in the cases without reliable stereo returns is discussed in [82].

With the associated distance map, the projected feature maps can be further combined or manipulated in a flexible way. We have shown an example in our experiment where two separate pairs of maps are obtained independently from both eye computers through their own stereo cameras. Shown in figure 5, the left eye is looking towards the right side to the robot (a)(b), while the right eye is looking towards left (c)(d), the data gathered for each individual eye is incomplete by essentially missing half of the information. The two maps are combined with the same heuristic that for each ground cell, features collected at a closer distance remains. The resulting feature map is shown in figure 6(a).

3.4.3 Data Labelling and Training

Once the feature map is obtained, we need to come up with a way to provide preferability labels and associate them to the training data, as previously discussed in section 3.2. However, there exists no obvious way of automated labelling process that matches exactly the preferability of the human teacher. Moreover, for the specific LAGR test, it is required that no human input for learning except the training log files provided.

We adopt the following heuristics based on the assumption that *at least* the terrain that the robot had driven on during training is preferable. This results in the procedure to label the ground within a distance D_{close} of the robot's trajectory as preferable (positive), while non-preferable (negative) data is obtained from the ground more than a distance D_{far} from the robot's trajectory. In order to minimize mislabelling, we conservatively set D_{close} to be quite small (1.2 meters) and D_{far} to be relatively large (4 meters). Any data



(a) (b) (c)

(d) (e)

Figure 6: Learning with ground projected feature map (a) Color feature map (b) Labelled training data: positive and negative examples are marked in green and red respectively (c) Encoded cost map classified with the learned model: preferable terrain are brighter (d) Color image from robot's view obtained via the stereo cameras (e) Classification based on learned color histogram model for "preferability".

points between D_{close} and D_{far} are considered ambiguous and are not used for training (figure 6(b)). Based on our design of the ground projected feature map and the distance map, the distance information for preferability labeling is immediately available. A color model is learned offline with the labeled training data.

The training procedure is a standard two-class classification problem. Depending on the feature space selected, several machine learning techniques can be used, such as Boosting, SVM, decision trees, etc. In our case, we learn a Bayes classifier for preferability with the $16 \times 16 \times 16$ RGB histogram. Let X be the appearance feature of RGB color (which is now the index number of a histogram bin), A be the value of preferability affordance. We calculate $p(x|A)$ for $A = 0, 1$ respectively, where $A = 1$ being preferable and $A = 0$ being non-preferable. We can then learn the following conditional probability of preferability:

$$\begin{aligned} p(A=1|x) &= \frac{p(x|A=1)p(A=1)}{p(x|A=1)p(A=1) + p(x|A=0)p(A=0)} \\ &= \frac{p(x|A=1)}{p(x|A=1) + p(x|A=0)\eta} \end{aligned} \tag{3}$$

where $\eta = \frac{p(A=0)}{p(A=1)}$ is ratio of the priors $p(A)$. In practise, it is unnecessary to threshold on this conditional probability to obtain a 0-1 binary classification, since the classification result is to be integrated into a cost map based additional channels of information (e.g. terrain height variation). Instead, we adopt the form of equation 3 to obtain a continuous cost function valued from 0 to 1, representing the degree of preferability: 1 being most preferable and 0 being least preferable. The η value is adjusted through experimentation with the planing and controlling module of the robot to obtain the best integrated performance. See figure 6(c) for applying this learned preferability function on the training feature map.

In testing, for each frame, based on the features calculated in the image (e.g. color), we can calculate the preferability of each pixel (figure 6(d)(e)). We then send the cost (or preferability) of each pixel and its relative location with respect to the robot to the control process, which is then integrated with other cost measures such as stereo determined obstacles. This is the same process as we did for traversability classification.

3.4.4 Experiment Results

Tests of the integrated system were performed at various test sites around the Atlanta, GA area. It should also be stressed that all algorithms presented here have been extensively tested and evaluated within the LAGR monthly competition format. For a complete description of the test setting and results, please refer to the official summary article by Jackel et. al. [36]. However, one nature of these competitions is the difficulty in objectively evaluating the relative merits of a system's individual components. Therefore we focus our discussion on the qualitative perception results, concentrating on the fact whether the preferability concept is learned from examples. More detailed discussion from a holistic system-wide viewpoint of our approach can be found in [82].

Test on mulch path One set of experiments of learning preferable terrain regions were tested at a vacant lot in Mableton, GA. The lot is primarily flat, but strewn with piles of landscaping waste (e.g. dead trees, bushes, and brush) creating a tangle of traversable "roads" among the large piles of obstacles, shown in Figure 7.

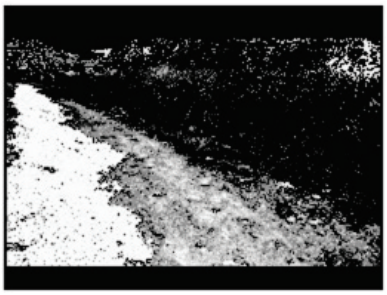
Figure 7: Test site for color-based learning of preferable regions.

Because of the tangled nature of the site's roads, there are potentially many acceptable routes between any two points at the site. For the test, start and goal points were chosen such that there existed a marginally traversable direct route to the goal and a longer but easily traversable route covered by cedar-wood mulch. The robot was allowed to run the course several times with out the benefit of learning, which established that the shorter but more difficult route was always chosen by the naive system. A human operator then guided the robot along another entirely separate mulch path, providing a training data set for the

color-based preferability classifier.

In the next round of tests, after offline preferability learning the robot was able to classify the mulch path as highly preferable. Figure 8(a)(b) compares a raw camera image of the mulch path and a classification of the same image. The mulch path is distinct in the classified image as an area of high preferability. The lower cost associated with this level of preferability caused the planner to lead the robot down the mulch path, and finally to the goal. The final cost map created by the robot is shown in figure 8(c). The mulch path is clearly visible as a region of low cost along the robot's trajectory.

(a)



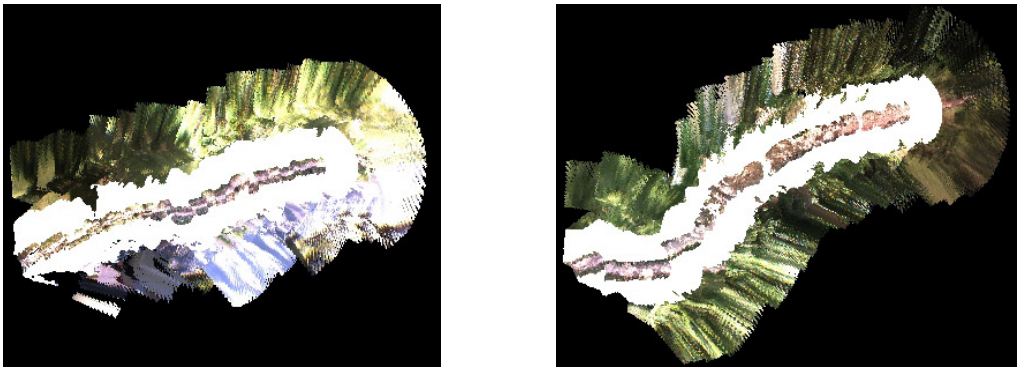
(b)

(c)

Figure 8: (a)(b) Example raw and classified camera images from the robot following the mulch path. In the classified image, lighter colored regions correspond to terrain learned to be preferable. (c) The final cost map from test on the mulch path. Lighter colored regions correspond to areas of lower cost (which, by extension, correspond to preferable terrain). The white line shows the robot's trajectory down the path. The location of the mulch path is clearly visible as a band of low cost along the robot's trajectory.

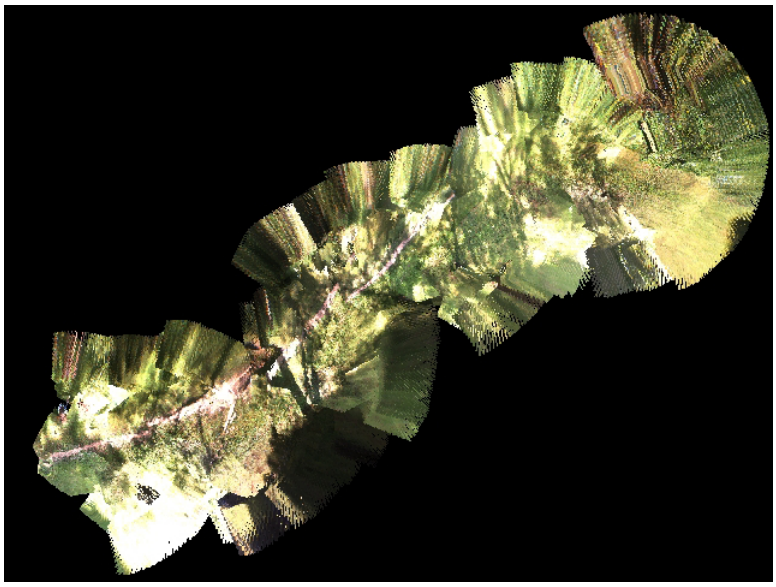
Test on muddy trial Another set of experiments were conducted in the Memorial park, Atlanta, Georgia. This site provide more challenging test environment with narrow muddy trial that is considered as preferable. The robot was trained by driving on two small segments of similar road. We show the color maps built during the training runs in figure 9. Only the positive training data (within 1.2 meters of the robot trajectory) and the negative training data (beyond 4 meters) are displayed.

In testing, the robot successfully detected the road and planned according to drive to the goal. The environment is visualized in figure 10(a) by the ground projected color map built during the testing course. The robot trajectory is illustrated by green line in figure 10(b), where the robot started from the place marked with a red dot on bottom of the map, and successfully travelled to the goal. We also show with gray scale image the raw preferability classification on the color map in figure 10(a) with the learned color model. It can be seen that the road is classified correctly from other terrain patches. The actual robot trajectory however is also affected by other behaviors in testing (such as stereo perceived obstacle avoidance), which explains the reason that the robot trajectory is not perfectly aligned with the detect road.

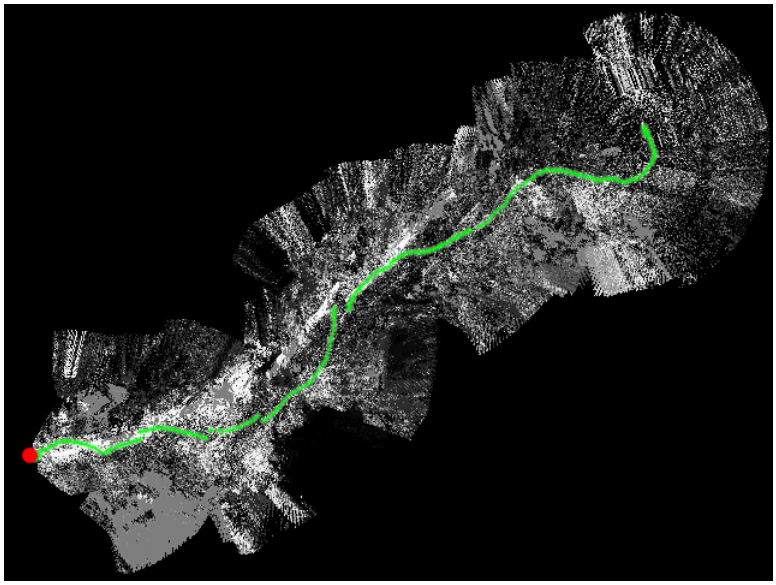


(a) (b)

Figure 9: The robot is driven on two small segments of similar road for training. Only the positive (road in the middle) and negative (terrain far from robot trajectory) training data are displayed while ambiguous data are discarded in training.



(a)



(b)

Figure 10: (a) The group projected color map built in testing, shown as a visualization of the test environment. (b) The image shows the robot trajectory (green line starting from the red dot) in testing where the robot was observed to attempt to drive on road whenever it could except the case when obstacles were detected on the road. The gray scale image is obtained by applying the learned color model (from the two training segment) on the test course color map, showing the raw preferability classification result.

3.5 Summary and Discussion

In this chapter, we discussed affordance learning with the direct perception approach. Two affordance learning tasks are presented as examples: traversability learning and preferability learning. For both tasks, terrain appearance is collected and labeled with affordance information for training. While the traversability information is collected by the robot interacting with its environment autonomously, the preferability information has to be provided by a human trainer by definition. Different learning algorithms, both online and offline, are presented. We demonstrate that a variety of robotics tasks can be casted as an affordance learning problem, and also illustrate the validity of learning a direct affordances perception system for challenging tasks in practice.

A limitation we observed is that direct perception learning requires a large amount of training data, because affordance in general is not a simple function of appearance. Direct perception can be characterized as learning at the object *instance level*, because a different instance (e.g. a patch of grass) of the same object category (e.g. grass) is treated as a separate training instance, completely independent from other training data that may come from the same object category. This contradicts the general observation that "although interaction with the world takes place at the level of individual objects, much reasoning takes place at the level of categories" [74].

As a result, direct perception is not an "adaptive" learning mechanism because new affordance learning does not build on the existing knowledge base of the affordances that are already learned, even if the new affordances may be highly correlated to the pre-learned ones. For example, if the robot were to learn the slippery affordance, and we further assume that in general the grass is slippery but the road is not, then it would be beneficial for the robot to generalize from its learning of the traversable and preferable affordances, which may not be possible with the direct perception approach.

In fact, because the objects in the environment are categorical with the physical properties of each object category being constrained, it should be the case that (at least some) affordances are dependent. Direct perception, as we demonstrated in this chapter, does

not make use of such categorical structure of the objects. Therefore, learning new affordances independently of the existing affordances with direct perception tend to require a large amount of data – this procedure with massive data requirement and little leverage of existing knowledge is at best suboptimal if not completely unacceptable.

An alternative way of learning is to acknowledge the categorical structure of the objects, by constructing a *shared intermediate representation of object categories* among the affordances. This categorical (or categorization) representation models the appearance distributions of the object categories and is shared among multiple affordances. This categorization representation will not only enable inference among multiple affordances through the category of an object, but also reduce the data requirement in new affordance learning because more general assertions of classification rules can be learned at the object category level instead of the instance level. In the next chapter, we provide a formal model to analyze how affordances are defined and introduce the way to incorporate object categorization as a means for information sharing in affordance learning.

CHAPTER IV

A GENERAL MODEL FOR OBJECT CATEGORIZATION AND AFFORDANCE PERCEPTION

4.1 The Category-Affordance Perception (CAP) Model

My thesis is that categorization can be used as an intermediate representation to improve affordance learning. The benefit in comparison to direct perception is the ability to use less data to learn affordance models. In particular, the reduction of data requirement can be significant in the case of incremental learning of new affordances.

Figure 11 presents our model of how affordance is defined and perceived by an agent. The object of interest is represented by its "physical properties" – in the broad extension – which includes the material, thickness, elasticity, and others that may or may not be easily measured. *Affordance is modeled as a deterministic function of the physical properties of the object and the agent.* In other words, the affordance is jointly determined by the physical properties of the object and the agent unambiguously and consistently – same physical properties always lead to the same affordance value. On the other hand, the object's appearance is a reflection of its physical properties through the agent's sensors. Likewise, the word "appearance" has its broad extension that includes any sensor measurement, such as color imagery, 3D geometry, sound and even laser scan profiles.[1] Because of sensor noise, appearance is a probabilistic function of the object's physical properties, the agent's sensors and the viewing conditions, i.e. they can only be determined in distribution.[2]

Defined as an action possibility that an object "affords" to the agent, an affordance determines the outcome of an experiment that the agent can (and might need to) perform on an object. In this sense, affordance can be "measured" through experiments. The outcome of a robot's attempts driving over an object, success or failure, provides a "measurement" of

[1] We assume that mental status is also a physical property. This work concentrates on non-living objects.
[2] Our definition of affordance is in the Gibsonian category instead of the representationist [11].

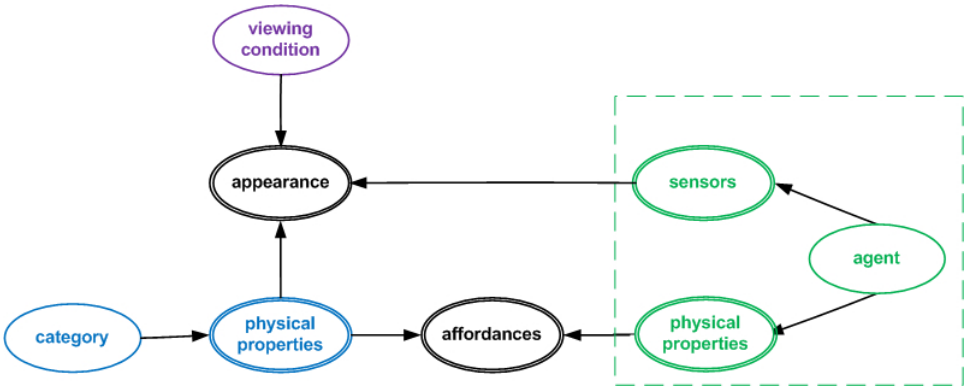


Figure 11: The category-affordance perception (CAP) model assumes that affordances are defined by the physical properties of both the object and the agent. The appearance of the object perceived by the robot is a (probabilistic) function of its physical property, the robot's sensors and viewing conditions.

the object's traversability affordance. In affordance perception, the agent makes a prediction of the affordance value based on the appearance information collected by its sensors. As mentioned earlier, Gibson [26] argued that the perception of affordance is a direct process where the agent forms a direct mapping from appearance to affordance prediction, where no intermediate representation is necessary.

Since the objects are categorical (as we discussed in chapter 1) the object category determines the distribution of physical properties. The goal of this work is to evaluate whether an intermediate representation of categorization can improve affordance prediction, and whether some categorizations are better than others in affordance prediction.

The CAP model presented can be used to model the affordance perception from multiple potentially different robot agents. To simplify the exposition we note that in the case of a fixed robot to consider, the model can be reduced by removing the agent specific elements of the model in figure 11. We show this fixed agent CAP model in figure 12(a), explicitly representing multiple affordances in multiple nodes.

4.2 A Probabilistic Characterization of the CAP Model

Illustrated in figure 12, let C denote category, X denote appearance, Y denote physical properties and A denote affordances (A^k being the k^{th} affordance), the CAP model specifies

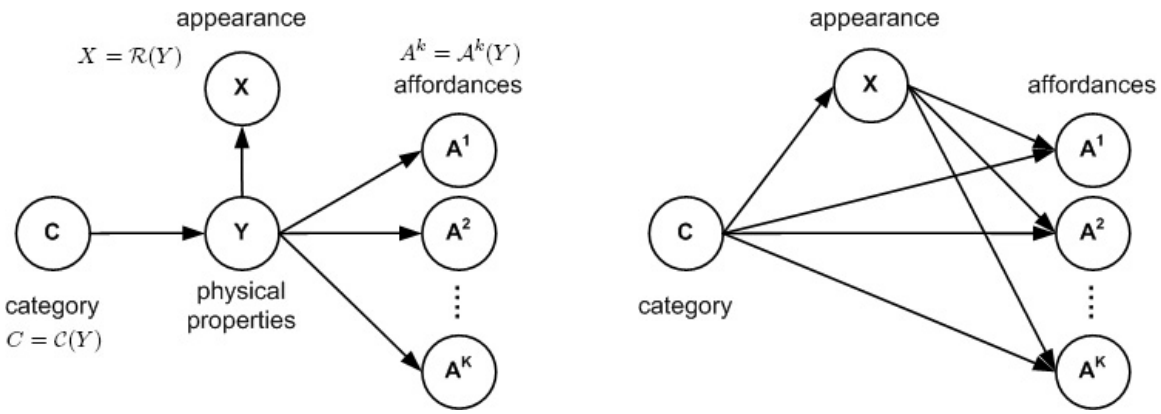


Figure 12: (a) CAP model for a fixed agent, the rendering, affordance and categorization functions are depicted; (b) A computational model for category-affordance learning without modeling the physical property Y, obtained by factorizing over Y and removed the dependencies among A^k.

the following joint density:

$$\begin{aligned} P(C, X, Y, A) &= P(C)P(Y|C)P(X|Y)P(A|Y) \\ &= P(C)P(Y|C)P(X|Y)\prod_k P(A^k|Y) \end{aligned} \tag{4}$$

Among these variables, C is discrete, representing one of the N object categories. X is a multidimensional continuous variable, which include both the raw measurements and the computed features. A is a K dimensional binary vector that each dimension represents one affordances. Finally we assume that Y is a quantitative representation of the object, which can potentially include any possible measure or description that describes the object and distinguishes it from others. Our model assumes the following relationships between these variables:

$$\textit{Rendering function:} \qquad X = \mathcal{R}(Y) \tag{5}$$

$$\textit{Affordance function:} \qquad A^k = \mathcal{A}^k(Y) \tag{6}$$

where the *rendering function* $\mathcal{R}$ specifies how the appearance is generated from the physical properties given a fixed set of sensors – it is probabilistic in the presence of measurement noise.[3] The *affordance function* as we defined previously now only depend on the physical

[3]Not being modeled explicitly, the view condition also contributes to the stochastic component of the rendering function.

Table 1: Variable definition of CAP model

Notation	Variable	Dimension	Comments
Y	Physical Properties	unknown, high	$P(Y\|C)$ specifies Y
X	Appearance	known, high	$X = \mathcal{R}(Y)$, probabilistic
C	Category	discrete	$C = \mathcal{C}(Y)$, deterministic
A	Affordance	binary	$A^k = \mathcal{A}^k(Y)$, deterministic

properties of the object given that the agent is fixed. We assume it to be deterministic without any probabilistic component. The notations are summarized in table 1.

Moreover, if we assume "unambiguous categorization" in the sense that given all the physical properties Y the category C of the object can be uniquely and consistently defined, we have the following *categorization function*:

$$\textit{Categorization function:} \qquad C = \mathcal{C}(Y) \tag{7}$$

This is indeed a classification function of the categories based on physical properties of an object. Compared with the model where we had specify a directed link pointing from C to Y, the categorization function can be thought of as a link from Y to C. Here we briefly mention that the CAP model specifies a "generative" model for C and Y while the categorization function is a "discriminative" model. Following discussion in [42], the major distinction is from the fact that a generative model specifies the class-conditional density as a characteristic function of the object category; while the discriminative model does not bother to model $P(Y)$ because the goal is to be able to do classification and as such Y is *always* provided as evidence. We discuss in detail the notion of generative versus discriminative both from a modeling point of view and a learning point of view later in section 5.1.

4.3 Three Approaches for Affordance Prediction

From the CAP model, if Y is given, other variables (X, C, A) are conditionally independent and can also be determined (though probabilistically for X). Ideally, if an inverse rendering function $\mathcal{R}^{-1}$ can be found, then affordance prediction and category recognition from appearance X can start from mapping X to Y. However, in practise the most difficult

variable to model is indeed the physical property Y – in general we do not know the number of dimensions and their extension (e.g. consider a physical property vector that defines traversability).

Therefore, it comes natural to model (X, C, A), by factorizing Y out from the joint density in equation 4 of the CAP model. We also assume conditional independence of the affordances given other variables (X, C), otherwise the attempt of modeling a fully connected graph of all possible affordances would be intractable. We refer to this computational model as the Category-Affordance (CA) model (see figure 12).

Different approximations or assumptions about the dependency results in several different approaches of affordance learning. We discuss three approaches: the direct perception approach, the category-affordance chain (CA-chain) model and the category-affordance full (CA-full) model. For each of the three approaches, we present our view from two different aspects. We start from a probabilistic graphical model point of view, to model the statistical dependencies of the random variables (X, C, A), without considering Y. Then we show how the approach can be derived from the CAP model and what approximations has been made.

4.3.1 The Direct Perception (DP) Approach

We begin with discussing the DP approach with respect to the CAP model, where we also examine the probabilistic approach for affordance prediction and its rationale. The goal of affordance perception is to predict the value of an affordance A^k given an observation of the appearance X. The DP approach aims at learning a direct mapping from X to A^k. Although Gibsonian DP theory argues for no intermediate feature representation or inference [26, 87], our notion of direct perception is rather general. Even if the feature representations are computed from appearance, it is still considered direct as long as the affordance prediction is based on this direct mapping (in contrast to a categorical approach for example). For one affordance, this implies:

$$P(X, A^k) = P(X)P(A^k|X) \tag{8}$$

Because X is always observed, DP does not need to model $P(X)$, and the only part

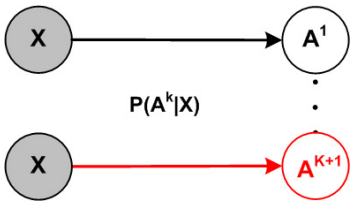


Figure 13: Graphical model of the direct perception (DP) approach. Red link shows the part of the model that needs to be learned for a new affordance A^{K+1}.

to learn is the conditional density of $P(A^k|X)$. In the DP model, multiple affordances are treated independently, i.e. the mapping from X to A^k are learned independently for each k^{th} affordance. This reflects a model of the following form by extending from equation 8:

$$\textit{DP Model:} \quad P(X, A) = P(X)P(A|X) = P(X)\prod_k P(A^k|X) \tag{9}$$

This inference is notionally depicted in figure 13. To see how this connects to the CAP model, we note that it can be derived from equation 4 with the following approximation:

$$\textit{DP Approximation:} \quad P(\mathcal{A}^k(Y)|Y) = P(\mathcal{A}^k(Y)|Y, \mathcal{R}(Y)) \;\approx\; P(\mathcal{A}^k(Y)|\mathcal{R}(Y)) \tag{10}$$

For a brief derivation, we note that this approximation implies:

$$P(A^k|Y) \triangleq P(\mathcal{A}^k(Y)|Y) \;\approx\; P(\mathcal{A}^k(Y)|\mathcal{R}(Y)) \triangleq P(A^k|X) \tag{11}$$

Substitute this into equation 4 of the CAP model, we have:

$$P(C, X, Y, A) = P(C, X, Y)\prod_k P(A^k|Y) \;\approx\; P(C, X, Y)\prod_k P(A^k|X) \tag{12}$$

Integrate both sides of the equation over C, Y, we obtain the DP model as in equation 9. This derivation is more intuitive and we also provide a mathematically rigorous derivation in the appendix A.

A Different View of the DP Approach: Beyond this discussion, we provide another useful way of looking at the DP approximation. Consider the case where Y can be observed, then all we need to learn for affordance prediction is the $\mathcal{A}^k$ function. The prediction of A^k based on observation $Y = y$ (denoted as $A^k|_{Y=y}$) is:

$$f(y) \;=\; A^k|_{Y=y} \;=\; \mathcal{A}^k(y) \tag{13}$$

In reality, what we observe is not Y but the result from the rendering function $\mathcal{R}(Y)$ or X, therefore in general we can't guarantee that a function mapping $f(X)$ be found:

$$f(x) \;=\; A^k|_{X=x} \;\equiv\; A^k|_{\mathcal{R}(Y)=x} \tag{14}$$

Nevertheless, the conditional expectation still exists:

$$\begin{aligned} \mathbf{E}[A^k|X=x] &= \int \mathcal{A}^k(y) p(Y=y|X=x)\; dy \\ &= \int \mathcal{A}^k(y) p(Y=y|\mathcal{R}(Y)=x)\; dy \end{aligned} \tag{15}$$

This conditional expectation is our best estimation of A^k. To summarize, the prediction $A^k|_Y$ can be made to be exact, but for the prediction of $A^k|_X$ (i.e. $A^k|_{\mathcal{R}(Y)}$), we use the conditional expectation $\mathbf{E}[A^k|X]$ as an approximation – this is the DP approximation we presented in equation 10. As previously discussed in section 3.2, we say that the appearance feature X is *sufficient* for affordance prediction if the conditional expectation is a good estimate, in the sense that the *expected error* is less than a pre-specified level η:

$$\mathbf{E}\left[\;\left|A^k - \mathbf{E}[A^k|X]\right|\;\right] < \eta \tag{16}$$

It is worth noting that the use of the conditional expectation as the estimator is equivalent to the probabilistic approach we discussed, in which the conditional probability $P(A^k = 1|X = x)$ is estimated for prediction. This is because the variable A^k is binary with value 0 or 1, which implies the following equivalence:

$$\mathbf{E}[A^k|X=x] \;\equiv\; P(A^k=1|X=x) \tag{17}$$

Pros and Cons of the DP approach: Finally we discuss the pros and cons of the DP approach. The advantage is its directness: given a task of predicting a binary variable A^k from input X, all that needs to be learned is the mapping function. It is in fact a binary classification function or classifier – a widely studied topic in machine learning. Also DP makes the least assumption about the structure of the environment and hence can be direct applicable even with the lack of domain knowledge.

However, the disadvantage is that DP approach does not provide much knowledge sharing among learning multiple affordances. Consider the case of learning a new affordance on

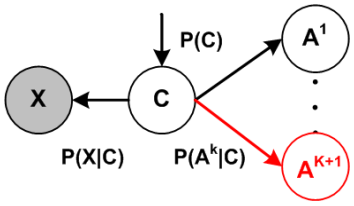


Figure 14: Graphical model of the category-affordance chain approach. Red link shows the part of the model that needs to be learned for a new affordance A^{K+1}.

top of a knowledge base of several affordances previously learned, probably the only thing to share is by which features to look at. The previous training data, without the new affordance value labeled, can hardly be useful to the new affordance learning. This weakness of the approach stems from the fact that independence affordance learning in DP does not make use of any categorical structure present in the world.

4.3.2 The Category-Affordance Chain Model

The category-affordance models (both the chain and the full version) takes advantage of a categorization of the objects. Category recognition has always been considered an important component in scene understanding in computer vision. The implication is that by recognizing the objects, we would be able to predict its properties and functionality, and ultimately how it might interact with the perceptual agent or other objects. In other words, recognizing the category of an object helps to predict its affordance.

We adopt this notion in the affordance prediction setting, such that category is an intermediate inference step in affordance prediction, and once the category of an object is known, the affordance can be predicted as well as possible. The inference process is depicted in figure 14 and the graphical model can be specified as:

$$\textit{CA-chain Model:} \quad P(X,C,A) = P(C)P(X|C)P(A|C) = P(C)P(X|C)\prod_k P(A^k|C) \quad (18)$$

By comparing with the CAP model in equation 4, we see that the following approximation has been made:

$$\textit{CA-chain Approximation:} \quad P(\mathcal{A}^k(Y)|Y) = P(\mathcal{A}^k(Y)|Y,\mathcal{C}(Y)) \;\approx\; P(\mathcal{A}^k(Y)|\mathcal{C}(Y)) \quad (19)$$

An intuitive derivation is given in the same way as we did for DP approach[4]:

$$P(A^k|Y) \triangleq P(\mathcal{A}^k(Y)|Y) \;\approx\; P(\mathcal{A}^k(Y)|\mathcal{C}(Y)) \triangleq P(A^k|C) \tag{20}$$

Substitute this into equation 4 of the CAP model, we have:

$$P(C,X,Y,A) = P(C,X,Y)\prod_k P(A^k|Y) \;\approx\; P(C,X,Y)\prod_k P(A^k|C) \tag{21}$$

Integrate both sides of the equation over Y, we obtain the CA-chain model as in equation 18.

We also refer to the CA-chain approximation as the the *sufficiency assumption of categorical affordance prediction*, because the approximation implies that knowing the object category is sufficient to predict the affordance, and no further appearance information will refine the prediction, i.e. $\mathbf{E}[A^k|C]$ is the best estimator of affordance. This assumption does not suggest that category determines affordance, but rather the affordance distribution is fixed given the category, independent of the appearance X.

As aforementioned, although the appearance to category and then affordance chain inference scheme directly leverages our ability to learn object categories, the validity of this approximation is related to the general question as why object instances are grouped together into categories and what defines a (good) category. In their paper about category utility [27], Gluck and Corter argues that the categorization process attempts to maximize the predictive power that predicts feature value from category labels. In affordance prediction for a robot agent, we assume that a useful categorization should enable the robot to make relatively accurate predictions about affordances once the category is recognized. Therefore the sufficiency assumption has to partially hold for at least some of the affordances. Furthermore, for the outlier category-affordance pairs when this assumption does not hold, we can leverage this model by learning a category specific affordance classifier only within the outlier categories. This extension is the CA-full approach presented in the following section.

[4]Mathematically rigorous derivations for the CA-chain/full models are similar to that given in the DP approach and are left out to avoid redundancy.

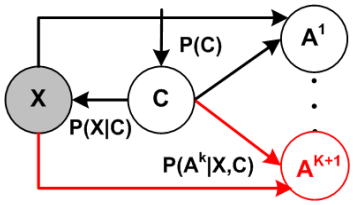


Figure 15: Graphical model of the category-affordance full approach. Red link shows the part of the model that needs to be learned for a new affordance A^{K+1}.

We also note that a key property of this categorical approach is that affordance prediction for multiple affordances share the same category structure (the $X - C$ link). Although affordances are conditionally independent of each other given the category value, they are *not* independent if the category is unobserved. This connection of affordances through category permits information about one affordance to provide constraint in making an inference about another. This is a major difference when compared with a direct perception approach. We will exploit this structure later in considering this information "sharing" through categories, which enables efficient learning of a new affordance from prior categorization knowledge, even with very limited amount of data.

4.3.3 The Category-Affordance Full Model

The previous section discuss the possibility of directly leveraging the category learning capability to affordance learning by adding a link between category and affordance based on the sufficiency assumption. If the sufficiency assumption holds in general, then the chain model for category-affordance prediction can be efficient learned on top of a pre-learned categorization. However, in the cases when the assumption does not hold, we can still make use of the categorical structure to learn the category specific appearance to affordance mapping. The "full" version of the category-affordance model is:

$$\textit{CA-full Model:} \quad P(X,C,A) = P(C)P(X|C)P(A|C,X) = P(C)P(X|C)\prod_k P(A^k|C,X) \tag{22}$$

Compared with the chain version of the model, we see that they both have the same category-appearance part in the model, they differ in the process of affordance prediction after inferring the category labels. The chain version assumes that the category itself is

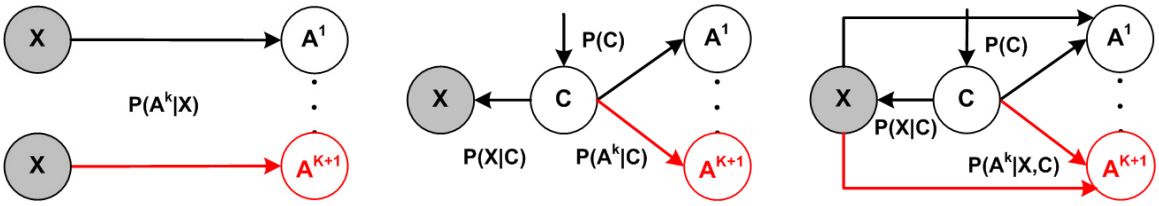


Figure 16: Three approaches for affordance learning: (a) direct perception (b) category-affordance chain (c) category-affordance full. Red link shows the part of the model that needs to be learned for a new affordance A^{K+1}.

sufficient to predict the affordance, while the full version requires additional information from the appearance. Similarly, coming from the CAP model, this category-affordance full model makes the following assumption:

$$\textit{CA-full Approximation:} \quad P(\mathcal{A}^k(Y)|Y) = P(\mathcal{A}^k(Y)|Y, \mathcal{R}(Y), \mathcal{C}(Y)) \approx P(\mathcal{A}^k(Y)|\mathcal{R}(Y), \mathcal{C}(Y)) \tag{23}$$

Notice that the category-affordance full model is a general case of the chain version. For categories when the sufficiency assumption holds, it reduces to the chain model; for outlier cases, it takes into account the appearance for affordance prediction. Compared with direct perception approach, the appearance to affordance mapping is learned on a per category basis.

4.4 Further Discussion

In this section, we compare the three affordance learning approaches presented earlier in his chapter: the direct perception approach, the chain and the full version of the category-affordance approach. Their graphical models are presented in figure 16 for comparison. We'll first discuss the difference of these approaches on whether to model the category or not. Then we summarize and compare the assumptions made in the 3 approaches.

4.4.1 The Categories: To Model or Not To

The DP approach models the direct dependency between appearance X and affordance A, while the CA-chain and CA-full approaches model an extra variable – category C. We note that not modeling C in the DP approach does not deny the existence of an categorical

structure. In other words, nonexistence of an categorical structure is not an assumption made by the DP approach, but rather it's a choice made on what variables matter and deserve modeling. Since the task is to predict A given X, learning a direct model of the probabilistic dependency of only these two variables is a reasonable approach that directly aims at the goal. The assumption that we made for DP is that the learning of different affordances are separable, i.e. they are conditional independent given X. Making this assumption allows the learning to focus on an individual affordance at a time. This assumption not only makes the learning in a more tractable scale, but also allows incremental learning of a new affordance (otherwise it may be required to relearn the entire appearance-to-affordances model). It also enables a direct prediction of the affordance of interest in testing, without a tedious procedure to first infer about other affordances and marginalize them out. However, it does not allow training data X for different A^k to be shared – they are made separate as well.

Similar as in the DP approach, both CA models assumes conditional independence of the affordances – for tractability concern as well as a means to allow incremental learning. The rationale for modeling an additional variable C in the CA models is from the assumption that a categorical structure exists in the environment. In other words, objects can be clustered with some similarity measure – unknown but does exist – in the physical properties Y space. Therefore knowing the category provides some information on A and X. This information can be probabilistic, for example "a table is normally not movable" and "tables are more likely to be in white color than green". Mathematically speaking, the prior distribution on X and A is different from the posterior distribution when given C. Or in other words, the mutual information between C and X, or C and A is nonzero.

One would naturally ask, is the the existence of categorical structure a strong assumption made by the CA models? This question can be addressed from two aspects. First if the data are not categorical, then a method to infer a latent category variable and then do affordance prediction should not have any significant advantage than a direct appearance to affordance inference method. This is true because in general we cannot guarantee that any learning task can be made easier just by dividing the data into groups to perform

learning on each group. Second, while we know natural object categories do exist in real world, the question here is to what extent an explicit modeling of the categories helps affordance learning. In the worst case scenario, a good categorization may be impossible to learn or it may be true the category does not provide useful information of the affordance – therefore irrelevant for affordance learning. Our work is to evaluate two different level of assumptions about the categorical structure and their influence on affordance learning in a number of aspects including data size, prediction performance, and learning complexity. At least one advantage that we can see directly from the figures is that training data for different A^k together contributes to the learning of the $X - C$ component of the model. This information sharing will be very useful in learning a new affordance, where the $X - C$ link is used as a prior knowledge.

4.4.2 Comparison of Model Assumptions

Besides the conditional independence assumption of the affordances, the CA-full model does not makes any additional assumptions. In principle, we expect that the CA-full approach be no worse than the DP approach, because DP is a special case of the CA-full model with one single category. In contrast, the CA-chain approach makes a very strong assumption that the category C information is sufficient to predict affordances A. Comparing with the CA-full model, we see that the sufficiency assumption we discussed previously can be written as:

$$P(A^k|X, C) = P(A^k|C) \tag{24}$$

Based on this assumption, once the category C is given, the affordance A is known in distribution – a Bernoulli distribution by definition, where A has probability p being 1 and probability $1 - p$ being 0. We have discussed that the validity of this assumption depends on whether objects are categorized to facilitate affordance prediction. For best affordance prediction performance, we see that this p parameter should be close to 0 or 1, such that less confusion remains when C is known. This connects the categorization to the category utility concept in the context of affordance prediction.

The independence assumptions for the three approaches from a graphical model point

Table 2: Summary of the three affordance learning approaches. We provide two different views for the comparison: one as an approximations to the CAP model, the other from a Bayes Network view where the independence assumptions are listed.

	Approximation to CAP $P(A^k\|Y) \approx$	Sufficient Information Set	Affordance Learning Independence Assumption
DP	$P(A^k\|\mathcal{R}(Y))$	X	$A^k \perp A^j \mid X$
CA-chain	$P(A^k\|\mathcal{C}(Y))$	C	$A^k \perp A^j \mid C$ $A^k \perp X \mid C$
CA-full	$P(A^k\|\mathcal{R}(Y), \mathcal{C}(Y))$	(X, C)	$A^k \perp A^j \mid C, X$

of view are summarized in the rightmost (fourth) column in table 2. The second column summarizes the approximations to the CAP model being made in the three approaches in order to circumvent the difficulties of explicitly modeling the physical properties Y.

Following our discussion in section 4.3.1 for the DP case, the three approaches differ what *information set* $\mathcal{G}$ to assume in estimating the posterior probability of the affordance $P(A^k = 1|\mathcal{G})$, or equivalently estimating the conditional expectation of the affordance $\mathbf{E}[A^k|\mathcal{G}]$. This is summarized in the third column as well. Since A^k is a deterministic function of Y, but not X and/or C, all the three approaches uses the conditional expectation estimator (depending on different information sets) as an approximation. The affordance prediction performance in practise depends on how well these approximations hold in the real world.

CHAPTER V

THE LEARNING ALGORITHMS FOR THE CA MODEL

In this chapter, we discuss different methods of training the CA model. Roughly speaking, *generative training* aims at maximizing the joint log likelihood function computed from the joint density $P(A, X)$. *Discriminative training* aims at maximizing the conditional log likelihood function computed from the conditional density $P(A|X)$. The methods differ in the objective function maximized, but are the same model as characterized by the graphical dependencies and the parameter set. The distinction between these training methods will be made clear in the exposition and the examples.

For both generative and discriminative training, we explore an optimization procedure based on generalized EM (GEM) that treats unobserved category labels as the missing data. We show that the assumption of a *shared categorization representation* results in a decoupled learning of the categorization (or category-appearance) model $P(X|C)$ and the the appearance-to-affordance classifiers $P(A|X, C)$ of each object category. Since the categorization is "shared", training data for different affordances all contributes to the learning of this part of the model. The difference between generative and discriminative training of the CA model is whether this categorization in turn should be learned generatively or discriminatively. On the other hand, the appearance-to-affordance classifiers are learned independently *within* each object category in a divide-and-conquer fashion.

Alternatively, we also explore training with direct optimization of the objective log likelihood function, based on the subset gradient decent method which takes advantage of the categorical structure of the model parameters by searching iteratively in the "natural" subsets of model parameters.

Finally, we discuss a different treatment of the category labels provided in training as well as adopting the model for classification in different scenarios of observations.

5.1 Comparison between Generative and Discriminative in CA Models

In this section, we compare generative and discriminative approaches, both in the *model* design and in the *training* algorithms, within the context of the category-affordance models. The general task is to model the statistical dependency between observation and target label. Appearance X is the observation, while each affordance A is the target label for affordance learning and the category C is the target label for category learning.

5.1.1 Generative and Discriminative Models

For *generative models* the (conditional) distribution of the observation is explicitly modeled in the form of $P(X|A)$. To predict the target label from given observation (e.g. to predict affordance from appearance), the Bayes law is applied to calculate the posterior distribution $P(A|X)$. This indirect inference procedure for target label classification inspires the design of the *discriminative models*, which do not model the observation distribution – because it is always observed in a classification task – but instead chooses to model $P(A|X)$ directly [42]. This can also be regarded as learning a classification function from X to A. These discriminative models are also called the *classifiers*. For example, Gaussian mixture models (GMM), probabilistic latent semantic analysis (pLSA) [31] and latent dirichlet allocation (LDA) [8] are generative models, while boosted classifiers [22] and support vector machines (SVM) are discriminative models. For more extensive discussion on this topic and its application in object recognition, the reader is referred to [19, 32, 46, 61, 89, 35]; in the next section, we discuss the choices that we make in designing the CA models.

In the CA models, we model the the category-appearance dependency generatively, which can be seen as a model of the *extension* of the object categories. This provides flexibility in learning object categories, where new categories can be effectively modeled by learning their own appearance distributions, without affecting existing categories. Because the categorization model is designed as a *shared* representation among different affordances, it constrains that the distribution of appearance X not depending on different affordances A^k. Therefore, it is natural to model the affordance prediction from the both the category and the appearance (CA-full), or only the category (CA-chain), or otherwise only the

appearance (DP). Therefore, in the CA-full approach[1], affordances are modeled discriminatively as a classifier based on the input of category and appearance. Moreover, because the category is a discrete variable, these affordance classifiers can be regarded as indexed by the category label (and also the different affordances), hereby referred to as the *category specific affordance classifiers*.

5.1.2 Generative and Discriminative Training

With the same CA model factorization, it is possible to train it with either the generative approach or the discriminative approach. We denote the set of i.i.d. training data as D, containing both the weakly and strongly labeled data.

In generative training, the best parameter θ is chosen to maximize the following log likelihood (`LL`) function on the joint density $P(A, X)$:

$$\texttt{LL}(\theta; D) = \sum_n \log p(a_n, x_n|\theta) \tag{25}$$

On the other hand, in discriminative training it is the conditional log likelihood function (`CLL`) on $P(A|X)$ that is to be maximized:

$$\texttt{CLL}(\theta; D) = \sum_n \log p(a_n|x_n, \theta) \tag{26}$$

Therefore, the learned model parameters are different for generative training Θ_G and discriminative training Θ_D. It is in this regard that recently there has been discussion in the literature as to whether the term "discriminative training" is a misnomer. For example, [56, 46] suggest that there is only one correct way of training a model and that maximizing `CLL` indeed implies a different model. However, here we will use the term in the conventional way to make a distinction to the discriminative models aforementioned. Therefore, training a categorization, with the appearance distribution modeled by Gaussian mixtures, by maximizing the `CLL` function is referred to as the *discriminative training of a generative model*.

[1] We concentrate on the discussion of the CA-full approach, while CA-chain is a special case and DP does not make use of the categorization.

In our category-affordance models which factorize over the category C, the LL and CLL functions are expanded as follows:

$$\texttt{LL}(\theta; D) = \sum_n \log \sum_c p(a_n|c, x_n, \theta) p(c, x_n|\theta) \tag{27}$$

$$\texttt{CLL}(\theta; D) = \sum_n \log \sum_c p(a_n|c, x_n, \theta) p(c|x_n, \theta) \tag{28}$$

We see that for both equations 28 and 27, there is a summation term in the log function which makes it difficult to obtain a closed form solution. We will discuss in the following section how an EM framework results in *decoupled* learning in both generative and discriminative training, where the categorization and the affordance classifiers are learned separately at each EM iteration. But these two learning tasks are not independent because they both affect the posterior of the unobserved category variable. A further advantage of the decoupling is that standard training algorithms can be applied for both category learning and affordance classifier learning, making the CA model generally applicable to leverage the state-of-the-art research in object recognition and general classifier design. On the other hand, in a direct optimization approach where this decoupling is not offered, the search of classifier parameter space can be overwhelming, we could however, still utilize the categorical structure by search in a subset of parameters. We will also provide an example of subset gradient decent in training a CA-chain model with direct optimization.

Another observation is that these two quantities are closely related [23], because:

$$\texttt{LL}(\theta; D) \quad = \quad \texttt{CLL}(\theta; D) + \sum_n \log p(x_n|\theta) \tag{29}$$

This is expected because while LL aims at maximizing the joint likelihood of all the observed variables in the training data, CLL removes the contribution in the likelihood function coming from the unconditional density of the appearance, because these are always observed in testing. While the connection suggests that there is significant similarity in the procedure of maximizing both the CLL and the LL, differences still exist for the two different training goals. Normally the LL is easier to optimize than CLL as we shall see in the next few sections.

In the next sections, we will discuss different methods to maximize the LL and CLL functions in the CA model. We also discuss the model we adopted with $P(X|C)$ as a Gaussian mixture model and $P(A|C, X)$ as a logistic regression classifier.

5.2 Generative Training with Generalized EM

The summation inside the logarithm of quantities ranging over large order of magnitude causes instability in direct optimization of the LL and CLL functions. One way of avoiding this is to use Generalized EM to take the summation out of the logarithm. The following discussion follows the general description of GEM, such as in [41].

First, we treat the unobserved object category of each weakly labeled training data c_n as a missing variable. The complete data set is defined as $\tilde{D} = \{(x_n, a_n, \tilde{c}_n)\}$ including the missing variables. Therefore, the original LL function in equation 27 is called the "incomplete-data" LL, while the "complete-data" LL function is:

$$\texttt{LL}(\theta; \tilde{D}) = \sum_n \log p(a_n, \tilde{c}_n, x_n|\theta) = \sum_n p(\tilde{c}_n|\theta) p(x_n|\tilde{c}_n, \theta) p(a_n|\tilde{c}_n, x_n, \theta) \tag{30}$$

The EM algorithm consists of an E-step and an M-step, where the E-step calculates the expected value of the complete-data likelihood, given the current model $\theta^{(t)}$ and the observed data:

$$\begin{aligned} \textit{E-step:} \quad Q(\theta, \theta^{(t)}) &= E^{\theta^{(t)}}\left[\texttt{LL}(\theta; \tilde{D})|D\right] \\ &= E^{\theta^{(t)}}\left[\sum_n \log p(x_n, \tilde{c}_n, |\theta) p(a_n|\tilde{c}_n, x_n, \theta)\right] \\ &= \sum_n E^{\theta^{(t)}}\left[\log p(x_n, \tilde{c}_n|\theta) p(a_n|\tilde{c}_n, x_n, \theta)\right] \end{aligned} \tag{31}$$

where $E^{\theta^{(t)}}$ denotes the expectation being taken under the posterior probability on $\tilde{c}_n$ calculated from the current model $\theta^{(t)}$.

In the M-step, we find the best model parameter θ such that $Q(\theta, \theta^{(t)})$ is maximized, and treat it as the new hypothesis of model parameters: $\theta^{(t+1)} = \arg\max_\theta Q(\theta, \theta^{(t)})$. To expand equation 31 explicitly, we denote the posterior of category c for the n^{th} sample as ${q_{n,c}}^{(t)}$ and the overall posterior probability of category c as ${q_c}^{(t)}$ (the superscript (t) emphasize that it is based on model $\theta^{(t)}$):

$$\begin{aligned} {q_{n,c}}^{(t)} &\triangleq p(c|x_n, a_n, \theta^{(t)}) \propto p(c|x_n, \theta^{(t)}) p(a_n|c, x_n, \theta^{(t)}) \\ {q_c}^{(t)} &\triangleq \tfrac{1}{|D|} \textstyle\sum_n {q_{n,c}}^{(t)} \end{aligned} \tag{32}$$

Intuitively, ${q_{n,c}}^{(t)}$ is the current probabilistic "guess" of the object category for each training data, while ${q_c}^{(t)}$ is the estimation of the prior probability of an object category. Substituting this into equation 31 yields:

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= \sum_n \sum_c p(c|x_n, a_n, \theta^{(t)}) \log p(c|\theta) p(x_n|c, \theta) p(a_n|c, x_n, \theta) \\ &= \sum_n \sum_c {q_{n,c}}^{(t)} \log p(c|\theta) + \sum_n \sum_c {q_{n,c}}^{(t)} \log p(x_n|c, \theta) \\ &\quad + \sum_n \sum_c {q_{n,c}}^{(t)} \log p(a_n|c, x_n, \theta) \end{aligned} \tag{33}$$

5.2.1 Training Decoupling

The meaning of the E-step equation 33 can be manifested by dividing the model parameters to two subsets such that $\theta \triangleq (\phi, \psi)$ – ϕ being the category-appearance model and ψ being the appearance-to-affordance classifiers. It can be seen in equation 33 that the first and second terms involving $p(c)$ and $p(x_n|c)$ only depend on ϕ, while the third term involving $p(a_n|c, x_n)$ only depends on ψ:

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= \underbrace{\sum_n \sum_c {q_{n,c}}^{(t)} \log p(c|\phi) + \sum_n \sum_c {q_{n,c}}^{(t)} \log p(x_n|c, \phi)}_{Q_\phi} \\ &\quad + \underbrace{\sum_n \sum_c {q_{n,c}}^{(t)} \log p(a_n|c, x_n, \psi)}_{Q_\psi} \end{aligned} \tag{34}$$

Therefore, the training can be *decoupled* into a search for the best category-appearance model ϕ by maximizing the first two terms and a search for the best appearance-to-affordance classifiers ψ by maximizing the third term. This comes as an logical consequence from our design of the model, where the appearance is determined by the object categories but not the affordances. Therefore this same decoupling of ϕ and ψ will be present in discriminative training as well, discussed in section 5.3. For generative training in particular, the two terms in Q_ϕ can also be independently maximized where the first term relates to the category priors ϕ_π and the second relates to the conditionally independent appearance distributions ϕ_c for each category c.

A second decoupling can be derived by rearranging in equation 33 the order of summations over the category label c and data index n:

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= \underbrace{|D| \sum_c {q_c}^{(t)} \log p(c|\phi_\pi)}_{Q_{\phi_\pi}} + \sum_c \underbrace{\sum_n {q_{n,c}}^{(t)} \log p(x_n|c,\phi)}_{Q_{\phi_c}} \\ &\quad + \sum_c \underbrace{\sum_n {q_{n,c}}^{(t)} \log p(a_n|c,x_n,\psi)}_{Q_{\psi_c}} \end{aligned} \tag{35}$$

This implies that both the category-appearance model and the appearance-to-affordance classifiers can be learned *independently* for each object category, denoted as Q_{ϕ_c} and Q_{ψ_c}. Besides the fact that category priors sum up to 1, the property of an object category is specified by its *own* (conditional) appearance distribution, which is independent from other object categories. The appearance-to-affordance classification is also category specific, i.e. each object category defines its *own* decision functions for affordance inference from appearance. In other words, *each object category is **defined** by its appearance distribution together with its affordance properties (i.e. appearance-to-affordance mapping).*

In summary, the first decoupling results from introducing the object categorization as an intermediate representation between appearance and affordances; the second decoupling results from the existence of multiple object categories where both the appearance and the affordances of an object depends on its own categories but not other categories.

5.2.2 Model Optimization

Finally, defining $\theta \triangleq (\phi_\pi, \{\phi_c\}, \{\psi_c\})$, we arrive at the following optimization tasks for the M-step to obtain the model parameter for $t+1$ iteration:

$$\textit{M-step:} \quad {\phi_\pi}^{(t+1)} = \arg\max_{\phi_\pi} \sum_c {q_c}^{(t)} \log p(c|\phi_\pi) \tag{36}$$

$$\forall c \quad {\phi_c}^{(t+1)} = \arg\max_{\phi_c} \sum_n {q_{n,c}}^{(t)} \log p(x_n|c,\phi_c) \tag{37}$$

$$\forall c \quad {\psi_c}^{(t+1)} = \arg\max_{\psi_c} \sum_n {q_{n,c}}^{(t)} \log p(a_n|c,x_n,\psi_c) \tag{38}$$

We now discuss each of these three maximizations in detail with intuitive explanation of the category-affordance learning task as well as discussion about a particular example of

the model that we adopted. We aim at providing a detailed analysis here and part of this discussion is also applicable to discriminative training.

Category priors Equation 36 is maximizing the negative cross-entropy (i.e. minimizing the cross-entropy) between the posterior and the prior distribution of categories. We maximize it by setting the object category priors ϕ_π to the posterior category probability ${q_c}^{(t)}$.

Category appearance Equation 37 for *each individual category* is maximizing the log likelihood of the data which is weighted by the posterior. Most standard category-appearance models can be readily adopted and the learning algorithm only need to be modified slightly to incorporate weighted training data. For example, in our experiment where a Gaussian mixture model is used per object category, the standard EM learning for GMM is still applicable. Specifically, for each category c, the update calculation of mixture coefficient, mean and variance of each component is based on weighting the training data by its posterior ${q_{n,c}}^{(t)}$ – the probability of the training data being from this particular category.

Category-specific affordance classifiers For equation 38, the optimization is also performed for each individual category independently. The remaining difficulty comes from the fact that there are normally multiple affordances and each of them affect the value of $p(a_n|c, x_n, \psi_c)$. Therefore, for the same object category, the learning tasks of the multiple affordance classifiers $\psi_{c,k}$ are coupled, and can be overwhelming.

We make an approximation by assuming each observed affordance vector a_n in training only contains one affordance, which leads to the following optimization:

$$\forall c \ \forall k \quad {\psi_{c,k}}^{(t+1)} = \arg\max_{\psi_{c,k}} \sum_{n \in S_k} {q_{n,c}}^{(t)} \log p(a_n^k | c, x_n, \psi_{c,k}) \tag{39}$$

We refer to equation 39 as the *independent learning of category-specific affordance classifiers*, where S_k is the set of training data for k^{th} affordance. It is the union of all the training data such that the k^{th} affordance is observed in the vector a_n (i.e not in the form of $a_n^{\bar{k}}$):

$$S_k \triangleq \{n : \ a_n \neq a_n^{\bar{k}}\} \tag{40}$$

Notice that equation 39 is indeed equivalent to equation 38 if for each training data only one affordance is observed, otherwise it's an approximation. In fact, in practice we often do obtain data with very few affordances, because experiments for multiple affordances rarely returns the results at exactly the same time. Experimental support for this approximation is also provided in appendix D.

5.2.3 Classifier Learning

Next we discuss learning the affordance classifiers $p(a^k|c, x, \psi_{c,k})$ – classification functions from X to A^k for each fixed category-affordance pair (c, k). It is binary classification because A^k can only be valued 0 or 1, and the actual value of the probability can be treated as confidence weighted classification. The derivation is standard and hereby provided for the completeness of the discussion.

Using the same notation of classifier learning, we denote this appearance-to-affordance classifier as $f^{c,k} : X \to [0, 1]$:

$$f^{c,k}(x) \triangleq p(A^k = 1|c, x, \psi_{c,k}) \tag{41}$$

Noticing that a_n^k is a binary variable, we derive the following equivalence:

$$\begin{aligned} p(a_n^k|c, x_n, \psi_{c,k}) &= p(A^k = 1|c, x_n, \psi_{c,k})^{a_n^k} \left(1 - p(A^k = 1|c, x_n, \psi_{c,k})\right)^{1-a_n^k} \\ &= f^{c,k}(x)^{a_n^k} \left(1 - f^{c,k}(x)\right)^{1-a_n^k} \end{aligned} \tag{42}$$

Substituting this into equation 39, the function to maximize for each fixed pair of (c, k) is equivalent to:

$$\begin{aligned} & \sum_{n \in S_k} {q_{n,c}}^{(t)} \log \left[f^{c,k}(x)^{a_n^k} \left(1 - f^{c,k}(x)\right)^{1-a_n^k} \right] \\ = & \sum_{n \in S_k} {q_{n,c}}^{(t)} \left[a_n^k \log f^{c,k}(x) + (1 - a_n^k) \log \left(1 - f^{c,k}(x)\right) \right] \end{aligned} \tag{43}$$

The term in the summation is the (weighted) binomial log likelihood or also the (negative) cross-entropy between the indicator variable of affordance and the posterior probability of the affordance. The posteriors ${q_{n,c}}^{(t)}$ are weights of each training sample.

In our experiment with the logistic regression classifier, training is based on a weighted version of the *Iteratively Reweighted Least Squares* (IRLS) method, a common method for

the *Generalized Linear Models* (GLM) [53, 41]. In the generalized EM framework, the maximization step for each iteration can be relaxed to merely increasing the $Q(\theta, \theta^{(t)})$ function [97], therefore we only perform a few steps of IRLS at each EM iteration. The training algorithm is summarized in algorithm 1.

Besides logistic regression, most classifiers can be adopted as long as they can evaluate the conditional probability instead of providing the binary classification result only. This makes it possible to evaluate the posterior ${q_{n,c}}^{(t)}$ in the global EM procedure. A minor modification for training is to incorporate weighted training data. Boosting classifiers in particular are applicable and in fact a similar form of equation 43 is the exponential loss function, serving as an error bound of classification error, that is iteratively minimized boosting [22].

Algorithm 1 Generative Training in CA Model through Generalized EM

Objective Function: The LL equation 27:

$$\texttt{LL}(\theta; D) = \sum_n \log \sum_c p(a_n|c, x_n, \theta) p(c, x_n|\theta)$$

Define (for the t^{th} iteration) the posterior ${q_{n,c}}^{(t)}, {q_c}^{(t)}$ as in equation 32.
Define S_k be the training set of A^k as in equation 40: $S_k = \{n : \ k_n = k\}$
Initialize $\phi^{(0)}$, $\psi^{(0)}$.
for each iteration $t = 0, 1, 2, ...$ **do**
    Calculate ${q_{n,c}}^{(t)}, {q_c}^{(t)}$ from $\phi^{(t)}, \psi^{(t)}$
    {`Learn the category-appearance model`}
    Learn the category priors from equation 36:

$${\phi_\pi}^{(t+1)} = \arg\max_{\phi_\pi} \sum_c {q_c}^{(t)} \log p(c|\phi_\pi)$$

    for each category $c = 1, 2, ...N$ **do**
        Learn the category-conditional distribution of appearance as in equation 37:

$${\phi_c}^{(t+1)} = \arg\max_{\phi_c} \sum_n {q_{n,c}}^{(t)} \log p(x_n|c, \phi_c)$$

    end for
    {`Learn the affordance classifiers`}
    for each category $c = 1, 2, ...N$ **do**
        Learn the (category specific) appearance-to-affordance classifiers for each affordance by maximizing (or increasing) Q_ψ as in equation 48:

$$\forall c \quad {\psi_c}^{(t+1)} = \arg\max_{\psi_c} \sum_n {q_{n,c}}^{(t)} \log p(a_n|c, x_n, \psi_c)$$

        As an approximation, classifiers for different affordances can be learned independently as in equation 39:

$$\forall c \ \forall k \quad {\psi_{c,k}}^{(t+1)} = \arg\max_{\psi_{c,k}} \sum_{n \in S_k} {q_{n,c}}^{(t)} \log p(a_n^k|c, x_n, \psi_{c,k})$$

    end for
end for

5.3 Discriminative Training with Generalized EM

Part of the discussion about discriminative training follows the discussion in the previous section. Therefore, we concentrate less on the derivations, but more on the difference of discriminative training and the intuitive explanation of category-affordance learning.

In discriminative training, it is the CLL function (equation 28) that is to be maximized. Similarly treating c_n as the missing data, we have the following "complete-data" CLL function:

$$\texttt{CLL}(\theta; \tilde{D}) = \sum_n \log p(a_n, \tilde{c}_n | x_n, \theta) = \sum_n \log p(\tilde{c}_n | x_n \theta) p(a_n | \tilde{c}_n, x_n, \theta) \tag{44}$$

The E-step given the current model $\theta^{(t)}$ is:

$$\begin{aligned} \textit{E-step:} \quad Q(\theta, \theta^{(t)}) &= E^{\theta^{(t)}} \left[\texttt{CLL}(\theta; \tilde{D}) | D \right] \\ &= E^{\theta^{(t)}} \left[\sum_n \log p(\tilde{c}_n | x_n, \theta) p(a_n | \tilde{c}_n, x_n, \theta) \right] \\ &= \sum_n E^{\theta^{(t)}} \left[\log p(\tilde{c}_n | x_n, \theta) p(a_n | \tilde{c}_n, x_n, \theta) \right] \end{aligned} \tag{45}$$

Comparing this to equation 31, the *joint* category-affordance probability $p(\tilde{c}_n, x_n | \theta)$ in generative training is now replaced by the *conditional* category probability given the appearance. On the other hand, the $p(a_n | \tilde{c}_n, x_n, \theta)$ term remains the same. Substituting ${q_{n,c}}^{(t)}$ from equation 32, we have:

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= \sum_n \sum_c p(c | x_n, a_n, \theta^{(t)}) \log p(c | x_n, \theta) p(a_n | c, x_n, \theta) \\ &= \sum_n \sum_c {q_{n,c}}^{(t)} \log p(c | x_n, \theta) + \sum_n \sum_c {q_{n,c}}^{(t)} \log p(a_n | c, x_n, \theta) \\ &= \underbrace{\sum_n \sum_c {q_{n,c}}^{(t)} \log p(c | x_n, \phi)}_{Q_\phi} + \underbrace{\sum_n \sum_c {q_{n,c}}^{(t)} \log p(a_n | c, x_n, \psi)}_{Q_\psi} \end{aligned} \tag{46}$$

Decoupling in discriminative training Equation 46 suggests that discriminative training maintains the decoupling of training the category-appearance model and the appearance-to-affordance classifiers. Comparing this with equation 34 in the generative case, we see that the Q_ψ term is exactly the same, which can be maximized in the same way as we discussed (equations 38 to 43). The difference is indeed in learning the category-appearance models.

5.3.1 Discriminative Training of Categories with Subset Gradient Decent

While generative training of affordances results in generative training of the categorization model, *discriminative training of affordances lead to the discriminative training of the categorization.* The Q_ϕ term changes from maximizing the logarithm of the joint $p(x, c)$ to the logarithm of the conditional $p(c|x)$. In the M-step we have:

$$\textit{M-step:} \qquad \phi^{(t+1)} = \arg\max_{\phi} \sum_n \sum_c {q_{n,c}}^{(t)} \log p(c|x_n, \phi) \tag{47}$$

$$\forall c \quad {\psi_c}^{(t+1)} = \arg\max_{\psi_c} \sum_n {q_{n,c}}^{(t)} \log p(a_n|c, x_n, \psi_c) \tag{48}$$

Although classifier learning is the same as in generative training with equation 48 being identical to equation 38, the category-appearance learning is different. In equation 47, it is difficult to further simplify Q_ϕ as in generative training (e.g. into the priors and the conditional appearance distributions) because the conditional probability evaluation involves computing a normalization term in the denominator as follows:

$$p(c|x_n, \theta) = \frac{p(c|\theta)p(x_n|c, \theta)}{\sum'_c p(c'|\theta)p(x_n|c', \theta)} \tag{49}$$

Unlike the case in generative training, in discriminative training the appearance of one category is related to the appearance of other categories, because every category's conditional distribution contributes to the marginalization constant for evaluating the posterior in equation 49. Since the goal is to distinguish among all these categories, we arrive at a paradox that the categorization model does *not* need to correctly *describe* the appearance distributions of the categories, rather it must *classify* objects into their categories. For example, in a discriminative categorization model of the grass and road categories, it may state that "grass is blue and road is red", because grass does look more blue (than road) and road does look more red (than grass). The model is considered *correct* as long as there is no confusion or error in classification; the true colors of grass (green) and road (brown) do not have to be learned.

Optimization procedures For generative models in general, Lasserre et. al. [46] pointed out that it is difficult to maximize the conditional density with the EM algorithm and they

directly maximize CLL with the conjugate gradient method. This is true because we are using a generative model. Had we adopted a discriminative model from X to C, the form of equation 47 is exactly a cross-entropy minimization criteria in multi-class classification that can be directly learned. It is possible, for example, to use IRLS method for Generalized Linear Models with the softmax function [41].

We also point out that for some special generative models the CLL can be easy to optimize. For example, when the conditional density $P(X|C)$ is a single Gaussian (or other distribution in the exponential family), $P(C|X)$ can be directly learned with GLMs because it has an equivalent form as a discriminative model with the softmax decision function [42]. For our methods to be generally applicable to generative models, we chose to adopt a Gaussian Mixture model as an example in our discussion of the subset gradient decent optimization method that we develop.

Subset gradient decent The subset gradient decent method is based on the observation that the parameter space in ϕ can be divided into independent subsets each of which describes a single category, because the appearance X distribution is conditional independent given the category C. Note that this conditional independence is specified by our model and it does not imply that the tasks of learning each category's appearance are independent.

The optimization procedure can be described as follows: for each object category, we adjust its appearance model so that it is more distinguishable from all other categories. Notice that the separability among the rest of the categories are not affected at the same time. For an intuitive example, we can think of the object instances resides in a (high dimensional) space, at each time the learning algorithm pulls one category further from all other categories, keeping the pairwise distances among other categories unchanged.

In subset gradient decent, we search a subset of the parameters such that the overall data conditional likelihood in equation 47 increases. The parameter subset include both the category's appearance distribution and *all* the category priors – because they are constrained to sum up to 1. We arrive a the following procedure to maximize equation 47:

Algorithm 2 Subset gradient decent for categorization learning

1: {`do it iteratively`}

2: **for** for each iteration t = 1, 2, ... **do**

3: {`update one category at a time`}

4: **for** each category c = 1, 2, ... N **do**

5: $(\phi_\pi^*, \phi_c^*) = \arg\max_{(\phi_\pi, \phi_c)} \sum_c {q_{n,c}}^{(t)} \log p(c|x_n, \phi)$

6: update ϕ with (ϕ_π^*, ϕ_c^*)

7: **end for**

8: **end for**

In a Generalized EM setting, where increasing the function value of equation 47 is sufficient rather than maximizing it in each iteration, the maximization in subset gradient decent can be replaced by updating the parameter in its gradient direction. Limiting the search only in a subset of parameters results in a significant speed-up compared with other methods based on gradient of the full parameter set. The complete algorithm of discriminative training with EM is summarized in algorithm 3.

Algorithm 3 Discriminative Training in CA Model through Generalized EM

Objective Function: The CLL equation 28:

$$\texttt{CLL}(\theta; D) = \sum_n \log \sum_c p(a_n|c, x_n, \theta) p(c|x_n, \theta)$$

1: Define (for the t^{th} iteration) the posterior ${q_{n,c}}^{(t)} = p(c|x_n, a_n, \theta^{(t)})$ as in equation 32.
2: Define S_k be the training set of A^k as in equation 40: $S_k = \{n : \ k_n = k\}$
3: Initialize $\phi^{(0)}$, $\psi^{(0)}$.
4: **for** each iteration $t = 0, 1, 2, ...$ **do**
5: Calculate ${q_{n,c}}^{(t)}$ from $\phi^{(t)}, \psi^{(t)}$
6: {`Learn the category-appearance model`}
7: Learn the category-appearance parameters $\phi_{(t+1)}$ by maximizing (or increasing) Q_ϕ as in equation 47:

$$\phi^{(t+1)} = \arg\max_\phi \sum_n \sum_c {q_{n,c}}^{(t)} \log p(c|x_n, \phi)$$

This can be done using the subset gradient decent procedure in algorithm 2.

8: {`Learn the affordance classifiers` (same procedure as in algorithm 1)}
9: **for** each category $c = 1, 2, ...N$ **do**
10: Learn the (category specific) appearance-to-affordance classifiers for each affordance by maximizing (or increasing) Q_ψ as in equation 48:

$$\forall c \ \ {\psi_c}^{(t+1)} = \arg\max_{\psi_c} \sum_n {q_{n,c}}^{(t)} \log p(a_n|c, x_n, \psi_c)$$

As an approximation, classifiers for different affordances can be learned independently as in equation 39:

$$\forall c \ \forall k \ \ {\psi_{c,k}}^{(t+1)} = \arg\max_{\psi_{c,k}} \sum_{n \in S_k} {q_{n,c}}^{(t)} \log p(a_n^k|c, x_n, \psi_{c,k})$$

11: **end for**
12: **end for**

5.4 Training through Direct Optimization

A direct way to maximize the objective functions for both generative (LL) and discriminative training (CLL) is via search methods such as gradient decent or conjugate gradient. Starting from an initial guess of the parameters, we iteratively increase the function value until arriving at a local maximum. We note that the initialization is possible given the strongly labeled data with category labels. Therefore we concentrate on the discussion of the optimization given an initial guess.

While direct optimization is theoretically straightforward, in practice there are stability concerns. For generative training, we note that equation 27 involves calculating $\log \sum_c p(a_n|c, x_n)p(c, x_n)$ for each training sample. $p(a_n|c, x_n)$ is the posterior of the observed affordance given the category and appearance and, as a probability of a binary variable, its value is bounded between 0 and 1. But the value of the probability density function $p(c, x_n)$ can range from a large order of magnitude (e.g. from 10^{-30} to 10^{30} for Gaussian mixture models). Adopting direct optimization is unstable because of the cancellation error. For discriminative training (equation 27), we see that because the joint $p(x, c)$ is replaced by $p(c|x)$ in the logarithm, stability is not a concern in direct maximization of the CLL function. Since c is a discrete variable, the probability mass function $p(c|x)$ is bounded. We outline a direct optimization procedure based on subset gradient decent introduced in the discussion with EM-based training in section 5.3.

5.4.1 Direct Optimization with Subset Gradient Decent

First by examining the natural of the category-affordance model, we discuss the most natural way of dividing the parameters into subsets. In section 5.2.1, we have shown that (ϕ_π, ϕ_c) is a subset of parameters of object categories' appearance distributions that are shared among multiple affordances (and not dependent on any particular ones). Besides, $\psi_{c,k}$ is also a natural subset that defines a classifier for the k^{th} affordance in the domain of the c^{th} category. In EM approach, we discussed that a particular classifier $\psi_{c,k}$ does not depend on affordance classifiers of other object categories, but depends on other affordance classifiers of the same object category. This is because that the weight of the training data for $\psi_{c,k}$ is

affected by classifiers of other affordances. This motivates the approximation of independent learning of category-specific affordance classifiers discussed in section 5.2.2.

In the direct maximization approach, no such approximation is required because classifier learning is by direct search of the parameter space rather than using typical methods to learn a classifier on weighted data. At each step, we search in the direction as specified by the subset of the parameters.

To facilitate our discussion, we define the following notation:

$$g_{n,c}{}^{(t)} \triangleq p(c|x_n, \phi^{(t)}) \tag{50}$$

$$h_{n,c}{}^{(t)} \triangleq p(a_n|c, x_n, \psi_c{}^{(t)}) \tag{51}$$

where $g_{n,c}{}^{(t)}$ is the contribution of the categorization model to the `CLL` function, and $h_{n,c}{}^{(t)}$ comes from the appearance-to-affordance classifiers. The subset gradient decent can be seen to greedy maximize each of these components of the objective function in an iterative fashion.

The algorithm of direct optimization is summarized in algorithm 4. Line 14-20 describes the learning of each category specific affordance classifiers. We initialize the classifier parameters $\psi^{(t+1)}$ from last iteration. Then for each of the classifier $\psi_{c,k}$, we search for its optimal parameters to maximize `CLL` while keeping other classifier parameters and the appearance parameters constant. Each newly learned classifier updates $\psi^{(t+1)}$, which is immediately reflected in the objective function in equation 53 and hence affects the next classifier learning. The order of learning the classifiers can be randomized. Comparing equations 52 and 53 (in the algorithm figure) with equations 47 and 48 in the EM approach, we see the difference arises from the existence of "log of sum" evaluation in direct optimization. In EM, the logarithm of both the category conditional density and the affordance conditional density are weighted by the posterior density of the categories $g_{n,c}{}^{(t)}$, which makes it possible to adopt standard classifier learning methods. However, this is not possible with the direct optimization approach, we resort to maximizing the log of sum function directly through searching, which can be done by subset gradient decent.

Algorithm 4 Discriminative Training in CA Model through Direct Maximization

Objective Function: The CLL function in equation 28:

$$\texttt{CLL}\ (\theta; D) = \sum_n \log \sum_c p(a_n|c, x_n, \theta) p(c|x_n, \theta)$$

Define (for the t^{th} iteration) ${g_{n,c}}^{(t)}, {h_{n,c}}^{(t)}$ as in equations 50 and 51:

$$g_{n,c}{}^{(t)} \triangleq p(c|x_n, \phi^{(t)}),\ \ h_{n,c}{}^{(t)} \triangleq p(a_n|c, x_n, {\psi_c}^{(t)})$$

Define S_k be the training set of affordance A^k as in equation 40: $S_k \triangleq \{n :\ a_n \neq a_n^{\bar{k}}\}$

Initialize $\phi^{(0)}$, $\psi^{(0)}$.

for each iteration $t = 0, 1, 2, ...$ **do**

  Calculate ${h_{n,c}}^{(t)}$ from $\psi^{(t)}$.

  {`Learn the category-appearance model`}

  Begin search with $\phi \leftarrow \phi^{(t)}$.

  for each category c **do**

    Maximize `CLL` w.r.t. (ϕ_π, ϕ_c):

$$({\phi_\pi}^{(t+1)}, {\phi_c}^{(t+1)}) = \arg\max_{\phi} \sum_n \log \sum_c h_{n,c}^{(t)}\ p(c|x_n, \phi) \tag{52}$$

    Update ϕ with $({\phi_\pi}^{(t+1)}, {\phi_c}^{(t+1)})$.

  end for

  Calculate ${g_{n,c}}^{(t+1)}$ from $\phi^{(t+1)}$.

  {`Learn the affordance classifiers`}

  Begin search with $\psi \leftarrow \psi^{(t)}$.

  for each category $c = 1, 2, ...N$ **do**

    for each affordance $k = 1, 2, ...K$ **do**

      Maximize `CLL` w.r.t. $\psi_{c,k}$:

$${\psi_{c,k}}^{(t+1)} = \arg\max_{\psi_{c,k}} \sum_{n \in S_k} \log \sum_c {g_{n,c}}^{(t+1)}\ p(a_n|c, x_n, \psi) \tag{53}$$

      Update ψ with ${\psi_{c,k}}^{(t+1)}$.

    end for

  end for

end for

5.5 An Alternative Training Goal: Fixing the Category Labels

Another important point is the treatment of category labels C that are also provided by the strongly labeled data. So far the learning objectives have not considered category labels. They are used only in initialization where $P(X|C)$ can be learned given the training data with category labels (step 3 of algorithms 1-3). It is natural to consider a different learning objective that incorporates the provided C labels. Therefore, for the strongly labeled set D_S, the LL and CLL functions are now concerned with $P(A, C, X)$ and $P(A, C|X)$, equations 25-26 for D_S are modified to:

$$\texttt{LL}(\theta; D_S) = \sum_{n \in S} \log p(a_n, c_n, x_n|\theta) \tag{54}$$

$$\texttt{CLL}(\theta; D_S) = \sum_{n \in S} \log p(a_n, c_n|x_n, \theta) \tag{55}$$

The new objective function for the entire training set is the sum of equations 27-28 for weakly labeled set and equations 54-55 for strongly labeled set:

$$\texttt{LL}(\theta; D_W, D_S) = \sum_{n \in W} \log p(a_n, x_n|\theta) + \sum_{n \in S} \log p(a_n, c_n, x_n|\theta) \tag{56}$$

$$\texttt{CLL}(\theta; D_W, D_S) = \sum_{n \in W} \log p(a_n|x_n, \theta) + \sum_{n \in S} \log p(a_n, c_n|x_n, \theta) \tag{57}$$

This can be incorporated in the EM based algorithms 1,3, but setting the posterior distribution ${q_{n,c}}^{(t)}$ to be 1 for the given category and 0 otherwise:

$${q_{n,c}}^{(t)} = \mathbf{I}(c_n = c) \tag{58}$$

Taking LL as an example (CLL is similar), by our definition of ${q_{n,c}}^{(t)}$ extended to strongly labeled set, the E-step is exactly the same form as equation 31:

$$\begin{aligned} \textit{E-step:} \quad Q(\theta, \theta^{(t)}) &= E^{\theta^{(t)}}\left[\texttt{LL}(\theta; \tilde{D_W}, D_S)|D_W, D_S\right] \\ &= E^{\theta^{(t)}}\left[\sum_{n \in W} \log p(a_n, \tilde{c}, x_n|\theta) + \sum_{n \in S} \log p(a_n, c_n, x_n|\theta)\right] \\ &= \sum_{n \in W \cup S} {q_{n,c}}^{(t)} \log p(a_n, c, x_n|\theta) \end{aligned} \tag{59}$$

Therefore, for training generatively or discriminatively both the CA-chain and the CA-full models, we can learn with or without fixing the category posterior for strongly labeled set with the EM framework. Experimental comparisons are discussed in detail in the next chapter.

5.6 Classification with Learned CA Model

In testing, we are given an appearance $\hat{x}$ and the goal is to predict the affordance a^k. The can be done by choosing the binary value of the affordance such that the posterior probability $p(a^k|\hat{x})$ is maximized:

$$\begin{aligned} \hat{a}^k &= \arg\max_{a^k=0,1} p(a^k|\hat{x},\theta) \\ &= \arg\max_{a^k=0,1} \sum_c p(c|\hat{x},\theta)p(a^k|c,\hat{x},\theta) \end{aligned} \tag{60}$$

This is in the case when no affordance label is provided. In the case when some other affordances (not a^k, the one to be inferred) are given, the posterior on the category label is affected by the known affordances. This in turn changes the affordance prediction:

$$\begin{aligned} \hat{a}^k &= \arg\max_{a^k=0,1} p(a^k|\hat{x},\hat{a}^{\bar{k}},\theta) \\ &= \arg\max_{a^k=0,1} \sum_c p(c|\hat{x},\theta)p(\hat{a}^{\bar{k}}|c,\hat{x},\theta)p(a^k|c,\hat{x},\theta) \end{aligned} \tag{61}$$

This suggests that in testing, it is possible to measure other affordances to improve the prediction of the affordance in question. Consider the distinction between "expensive" affordances and "inexpensive" affordances, where the cost can be defined by amount of resource involved in measuring the affordance, the degree of obstructive/destructive impact on the environment (such as noise) or the robot itself. Therefore the capability of measuring inexpensive affordances to help predict expensive affordances can be useful in certain situations. This is considered another advantage of the categorical approach over the DP approach.

Another task is to predict a *set* of affordances at the same time, where the success requires getting *all* the affordance predictions correct and any single affordance prediction error results in failure. This is important because some task requires the correct prediction of a set of affordances at the same time, which is equivalently defining a new affordance on these (base) affordances. Leaving the detailed discussion for section 6.2, we now write the classification rule of an affordance group as follows:

$$\begin{aligned} \hat{a} &= \arg\max_{\forall k\ a^k=0,1} p(a|\hat{x},\theta) \\ &= \arg\max_{\forall k\ a^k=0,1} \sum_c p(c|\hat{x},\theta) \prod_k p(a^k|c,\hat{x},\theta) \end{aligned} \tag{62}$$

This computes which set of affordances values are most likely — most consistent with each other and the appearance observation. It is computationally costly because the probability of $p(a|\hat{x})$ needs to be evaluated for 2^K number of all possible affordance values. Therefore we use an approximation of predicting each affordance independently via equation 60. The approximation is valid if one of the categories has probability $p(c|\hat{x})$ close to 1. Note that the approach of combining independently predicted individual affordances for group affordance prediction is the only possible way for the DP approach, because affordances are learned independently. We will discussed later in chapter 6 that the independent learning in the DP approach results a higher error for group affordance prediction compared to the CA models.

5.7 Summary and Discussion

We summarize the discussion on learning with the CA model. The learning algorithms presented in this chapter are all based on the model factorization with the category-appearance (i.e. categorization) model as an intermediate representation.

Sharing the categorization among multiple affordances has three major advantages. The first is that training data for different affordances all contributes to the learning of this categorization model. Therefore learning multiple affordances are no longer independent tasks. This can be an advantage when there is limited training data for some particular affordance. Second is that the affordance training is decoupled to two tasks of categorization learning and affordance classifier learning given the category. Optimization for these two tasks are independent at each iteration in an EM framework. The third advantage is in the case of learning new affordances, where eliminating the need of learning the shared categorization model enables learning with scarce training data and less computation. We will demonstrate these advantages with experiments on real data in the next chapter.

For both generative and discriminative training, an EM approach can be adopted which treats the unobserved category label as the missing training data. Based on the posterior distribution of the missing category label, both the categorization model $P(X, C)$ and the category-specific affordance classifiers $P(A|X, C)$ are learned independently at each EM

iteration.

The major difference between generative and discriminative training lies in learning the categorization model. Generative training of affordances requires also generative training of object categorization, while discriminative training of affordances requires discriminative categorization learning. While in the generative case, the category-appearance model can be learned with standard techniques (such as EM for Gaussian mixtures), the discriminative category training with a generative model is more complicated. As discussed in section 5.3, this is because that the calculation conditional probability $P(c|x)$ involves calculating a normalization term – a summation that depends on the appearance models for all categories. However, the existence of a categorical structure insures that the model parameters can be divided into disjunctive subsets, each corresponding to a single object category. The parameter maximization search can then be performed on subsets of parameters iteratively, which is significantly faster than search in the entire (high) parameter space. We introduce a subset gradient decent technique for discriminative categorization learning and discuss with the Gaussian mixture model as an example.

On the other hand, the second learning task resulting from the decoupling, i.e. the learning of category-specific affordance classifiers, are identical for both generative and discriminative training. This part of learning is also facilitated by the presence of a categorical structure, as affordance classifiers are now learned within each object category and those of different categories are independent. With the approximation of independently learning different affordance classifiers for the same object category, classifier learning is further decomposed to learning a standard binary classifier for each category-affordance pair.

Unlike the EM framework, a direct optimization approach does not have the two decoupled learning tasks of category-appearance model and affordance classifier, but maximize the objective function through a searching procedure. In our experiment, we only apply this direct maximization approach to the CA-chain model, where the classifier is simplified as the conditional probability: $P(A|X,C) = P(A|C)$. We show that the subset gradient decent can also be applied because of the categorical structure.

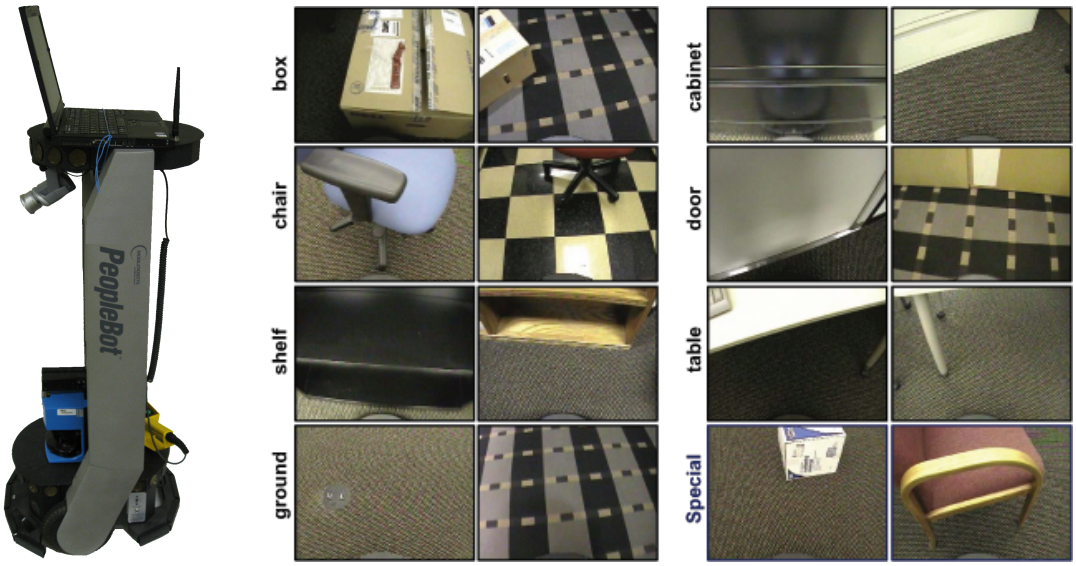
CHAPTER VI

EXPERIMENTS AND RESULTS

In this chapter, we present the experimental results with a robot working in the indoor office environment. In the first two experiments, we consider learning 3 affordances in a batch mode from both strongly labeled and weakly labeled training data. We study the affordance classification performance in two scenarios: learning with large training sets and with small training sets. Different training algorithms discussed in chapter 5 as well as the direct perception approach are compared. We show that although with large training sets the performance of DP and the CA approaches are comparable, the CA approaches outperforms in the case of limited amount of training data. The CA-chain model in particular, even though its discriminative power is limited by the assumption that categories determines affordances, is shown to achieve better than expected performance via re-organizing the category membership of objects — a process we referred to as *re-categorization*.

We also show that with the benefit of a categorical structure to connect different affordances, the CA approaches outperform the DP approach in terms of classifying a set of affordances at the same time, which is measured by the group affordance classification cost (defined in this chapter). Learning all affordances independently, the DP approach cannot avoid making inconsistent affordance predictions, such as predicting an object that is both traversable and movable, which can not exist.

Then we study the task of learning new affordances with the knowledge of a set of pre-learned affordances. We show that the CA approaches is able to generalize assertions about affordances for each object category, making it possible to learn the new affordances with very few training data. The DP approach, always required to learn new affordances independently "from scratch", suffers from a lack of training data. Moreover, the CA-full approach can go beyond the classification performance of the CA-chain approach by learning the category-specific affordance classifiers. Classification performance with training set size



(a) (b)

Figure 17: (a) The indoor robot (MobileRobots) used for our experiments. (b) The data set of 7 typical object categories indoors, with which 6 affordances are considered. The "special" objects refer to the cases that category does not determine affordance.

varying from small to large is studied in this experiment. We also provide guidelines of the typical procedures to learn new affordances.

The similar scenario is also applicable in learning with unbalance training data, in the case where some affordances have significantly smaller training sets than other affordances. In this case the categorization or category-appearance model learning is mostly affected by the affordances with abundant training data, whereas the affordance that lacks training data is learned similarly as learning a new affordance on an existing categorization model.

With affordance learning experimental results on large vs. small training data and on new affordance learning, we conclude that the category-affordance approaches, by incorporating knowledge of an object categorization, are advantageous over the direct perception approach in building an adaptive robot learner that requires less training data for affordance learning and adapts faster to new affordances. The categorical approach also helps to predict with more plausible combination of multiple affordance values that are consistent with each other.

6.1 Experiment Setting with an Indoor Robot

The following experiments were conducted on a standard indoor robot system, the Peoplebot [2]. As in figure 17, the Peoplebot is equipped with two wheels, both front and rear bumpers, and a point-tilt-zoom camera that is mounted at the height of approximately 1.2 meters. In our experiments, the camera position is fixed, pointing down to the ground to have a view of objects at a close distance in front of the robot.

6.1.1 Three Affordances: Traversable, Movable and Supportive

Seven most common object categories in an indoor office environment are considered in our experiments. The goal of our experiment is not to learn to recognize these categories, but to study whether utilizing an object categorization with these categories improves affordance learning over direct perception. Example images are shown in figure 17. We first study three affordances defined as follows:

A^1. *Traversable*: the traversable affordance is defined as whether the robot can successfully move forward, driving over the object — leaving it at its original location.

A^2. *Movable*: the movable affordance is defined as whether the robot can successfully push the object and change its location.

A^3. *Supportive*: the supportive affordance is defined as whether the robot can put down a small object (e.g. a cup of coffee) on top of the object.

Typically, boxes and chairs are movable, shelves and tables are supportive, and only the ground is traversable. We also relax these conditions by introducing non-typical objects, such as boxes and chairs that are not movable (see table 3 for summary). To keep the focus of category classification and affordance learning, we assume that affordances can be detected through experiments. The robot is manually driven in the lab to collect images of different objects. We define heuristics to parse the image logs collected to obtain affordance labeling. To simplify the parsing, we restrict the robot's driving patten with only straight forward/backward motion and point turns (turns with only an angular velocity).

For example, if the robot bumps into an obstacle and get stuck in forward motion, then any images collected back in time when the robot is within a maximum of 1.5 meter away from the current location is labeled as non-movable (and also non-traversable). The robot's 2D positional information is only used for log file parsing to obtain affordance labels, but not used as a feature for affordance learning.

Table 3: 7 object categories and their traversable, movable, supportive affordance properties. While most of the categories sufficiently determines affordance, we have special objects (non-movable boxes and chairs) as an exception. We later expand the set of affordances as in table 8.

Category	Traversable	Movable	Supportive
Box	No	Yes/No	No
Cabinet	No	No	No
Chair	No	Yes/No	No
Door	No	No	No
Shelf	No	No	Yes
Table	No	No	Yes
Ground	Yes	No	No

6.1.2 Implementation Details

In our experiment, we focus on affordance prediction based on a single camera image. This is different from our traversability and preferability learning work presented earlier where an explicit 3D occupancy mapping of the environment is performed. Note that sensors such as laser and stereo not only make it possible for 3D correspondence and thus occupancy mapping, but also provide both informative and accurate features such 3D depth maps. Unlike the imagery data, depth maps can be used to directly measure the size and shape of the object. We will come back to this issue later in chapter 8. With color camera as the only sensor, our task setting for affordance classification is to predict the affordances given a color image observation from the camera.

Otherwise stated, the following experiments are based on the same image features, same form of category-affordance models $P(X|C)$ and same form of affordance classifiers $P(A|X, C)$ explained shortly. However, it is important to note that our work is not about which specific feature or object model works the best on our data set. We are interested in comparing CA models and the DP approach in robot affordance learning, to study whether

object categorization are beneficial or necessary for this task. Other category models and features can indeed be incorporated in the same framework.

For each image, we extract its hue-saturation histogram of $18 \times 16 = 288$ bins, which is then reduced to a 10-D vector through Principle Component Analysis (PCA). This is concatenated by another PCA projected edge histogram feature vector. The edge histogram is calculated from a transformation similar to that of the census transform [100]. From each edge pixel on the edge image calculated by the Canny edge detectors, we map its 4 neighboring pixels to a 4 bit binary number — 1 if the pixel is also an edge pixel, and 0 otherwise. This binary number values from 0 to 15, we calculate a histogram of 15 bins from an image where the 8-bit transformed pixel value being 0 is discarded. This is again reduced to 5 dimensions using PCA. In summary, each image is represented by a 15-D vector containing both color and shape information. The dimension reduction makes it possible to learn a Gaussian mixture model for the object categories with a few amount of training data.

The rationale for choosing this representation is that color information is very helpful in indoor environments to recognize objects — for example, tables are normally painted in white or brown color. The edge histogram reflects the orientation distribution of the edges detected as well as corners and junction points in the edge map, providing shape information of the objects. The histogram feature is designed to be invariant to the location of the object in the image. This is different from the scene recognition work where the spatial configuration of an image is informative such that location-sensitive feature representation is pursued, such as in [49].

The category-affordance distribution $P(X|C)$ for each category is modeled as an Gaussian mixture in the feature space. We use 2 mixture components for each category, and fix the covariance matrix type to be diagonal. Therefore, there is 61 parameters for each category $P(X|C)$: 30 each for the 2 mixing component's mean and covariance and 1 for the mixing coefficient. The $P(X, C)$ model has $61 \times 7 + 6 = 433$ parameters, 6 for the prior of the category probability $P(C)$. For the affordance classifiers used in the CA-full model, we use logistic regression which can be learned with the weighted IRLS method. We

expect that within each category, the decision boundary of an affordance is in a simpler form than doing direct perception, therefore a logistic regression classifier probably suffice. Of course other complex classifiers can be used, for example a boosted ensemble classifier. For EM-based learning which involves evaluating the posterior distribution of the missing category label at each iteration, we require that the classifier can estimate the probability of $P(A|X, C)$ rather than produce a binary classification only.

The difference between the CA-chain model and the CA-full model is only in the form of the category-specific affordance classifiers. Note that for a particular pair of A^k and C with the probability $P(A^k = 1|C)$ being close to 0 or 1, learning the specific classifier $P(A^k|C, X)$ is not necessary. The CA-chain approach, however, approximates each affordance classifier with a binomial model specified by the probability $P(A^k = 1|C)$.

6.2 Evaluation and Algorithm Details

The first task we consider is to learn the 3 aforementioned affordances. We use equal number of images from each object category in the training data set. For each image, 1 or more affordance labels are provided, the category label can be known (e.g. strongly labeled) or unknown (weakly labeled). In testing, ground truth labels of the 3 affordances and the object category are provided for each testing image. This is used to evaluate the classifier performance in a number of tasks:

1. *Individual affordance classification:* the classification error of each individual affordance (e.g. traversability) measured separately. We also consider the average error for all the affordances, namely:

$$\xi_1(\theta; D) = \frac{1}{K\|D\|} \sum_{n \in D} \sum_{k=1}^{K} |\hat{a}_n^k - a_n^k| = \frac{1}{K\|D\|} \sum_{n \in D} \|\hat{\mathbf{a}}_n - \mathbf{a}_n\|_1 \tag{63}$$

   where $\mathbf{a}$ is the ground truth K-dimensional binary vector of the set of affordances being considered, and $\hat{\mathbf{a}}$ is the prediction based on the model θ and the appearance $\mathbf{x}$. The total error is equivalent to the L-1 norm of the affordance classification error vector $\hat{\mathbf{a}} - \mathbf{a}$.

2. *Group affordance classification:* another task is to detect a group of affordances for

each individual observation and error in any of the affordance classification results is assigned a cost of 1. This is relevant when the robot is performing a particular task which requires multiple affordances to be predicted correctly at the same time, while a misclassification of any one affordance results in the failure of execution. For example, the robot may seek objects that are non-movable and supportive to put down a small tool, which is then expected to be both stable and stationary. Similarly, this cost function of group affordance classification can be mathematically expressed as follows, which is in fact the ∞-norm of the error vector:

$$\xi_{\infty}(\theta; D) = \frac{1}{\|D\|} \sum_{n \in D} \max_{k=1,\ldots,K} |\hat{a}_n^k - a_n^k| = \frac{1}{\|D\|} \sum_{n \in D} \|\hat{\mathbf{a}}_n - \mathbf{a}_n\|_{\infty} \tag{64}$$

From the definition it is clear to see that:

$$\xi_{\infty}(\theta; D) \geq \xi_1(\theta; D) \tag{65}$$

For DP approach where each affordance classifier is learned independently, we can assume that errors on each individual affordance classification are independent. Denote each individual affordance classification error as ξ^k, the group affordance classification error ξ_{∞} defined in equation 64 can be approximately estimated as:

$$\xi_{\infty} \approx 1 - \prod_k (1 - \xi^k) \tag{66}$$

3. *Category recognition:* although our direct goal is affordance classification but not category recognition (and DP approach does not involve category recognition), it is worthwhile to compare the category recognition performance of different CA approaches. We are interested to see whether the category recognition performance is being pursued implicitly in order to best learn the affordances or whether on the contrary, they are negative correlated as in a trade-off. This evaluation is only for the category-affordance models but not for direct perception.

Next we discuss the 10 learning approaches which are to be compared in the experiments. They are based on different choices between CA-chain and CA-full model, generative or discriminative training, optimization with EM or direct search, and whether the category

Table 4: 10 learning algorithms for comparison, including direct perception (DP) and category-affordance (CA) models. The direct search algorithm is computational costly for CA-full model and not being considered.

Algorithm	Model	Objective Function to Maximize	Optimization Procedure	Fixing Category?
DP	DP	`CLL` (equation 26)	direct search	–
C1	CA-chain	`LL` (equation 56)	GEM	Yes
C2	CA-chain	`LL` (equation 27)	GEM	No
C3	CA-chain	`CLL` (equation 57)	GEM	Yes
C4	CA-chain	`CLL` (equation 28)	GEM	No
C5	CA-chain	`CLL` (equation 28)	direct search	No
F1	CA-full	`LL` (equation 56)	GEM	Yes
F2	CA-full	`LL` (equation 27)	GEM	No
F3	CA-full	`CLL` (equation 57)	GEM	Yes
F4	CA-full	`CLL` (equation 28)	GEM	No

labels are fixed in posterior calculation for EM (as described in section 5.5). The complete list of the 10 different training methods we consider can be found in table 4. For each method, with EM or direct search, an iterative optimization is involved. For each of the discriminative training methods (C3-5, F3-4), we use the learned model from C1 as an initialization. The training is continued for a large number of iterations. Then a post-training model selection step is involved to evaluate the training error with the model learned at each iteration and the one with the minimum average affordance classification error (equation 63) is selected. Fewer iterations are preferred to break the tie. To summarize, we maximize either the `LL` or `CLL` function in optimization, but the best model is selected based on the affordance classification error. The idea is that although `LL` and `CLL` can increase at each iteration, the training error does not improve and therefore it is reasonable to stop the optimization procedure to prevent over-fitting. Finally, this learning algorithm is performed multiple times for the same training data to select the learned model with the minimum training error, which is essentially random-start hill climbing.

6.3 Learning with Large Training Set

In the first experiment, we evaluate different training algorithms on a large training set as an empirical analysis of their asymptotical classification performance, which is difficult to reason theoretically. The training set contains 1400 images in total, 200 image for each object category. For each training image, all the 3 affordance labels are provided. Half of the images (700) are also provided with the category labels. The testing set contains about 6000 images, each labeled with the object category and the 3 affordance values. Although most categories do determine affordance, the fact that chairs and boxes can be either movable or non-movable provides a challenging situation for the CA-chain model.

The affordance classification performance are summarized in table 5 for the training and the testing data. In general, discriminative training (maximizing `CLL`) for both CA-chain and CA-full models results in smaller error compared with generative training (maximizing `LL`) in individual affordance classification, group affordance classification, and category recognition. "Generative training" for the CA-full model achieves comparable performance, because the logistic regression classifiers for $P(A|X, C)$ are in fact discriminatively learned. Therefore although theoretically maximize the `LL` function is considered generative training, it is a hybrid between generative training (i.e. Gaussian mixture learning with EM) and discriminative training (i.e. weighted IRLS to learn the logistic regression classifiers). It is also the reason why the CA-full model outperforms the chain model in every aspects too — a benefit of the increased model complexity. With discriminative training of maximizing `CLL`, the training error for group affordance classification (which is a upper bound for each individual affordance error) with the CA-full model (F3,F4) can be made lower than 1.0%. They also achieve the best testing error, with group affordance classification error of 5.2% (F3) and 5.7% (F4).

With large amount of training data, whether we fix the category labels in training or not does not affect affordance classification performance, although the category recognition performance is better in the former case, which is expected because the learning objective is to favor correct category recognition (see the comparison between C1&C2, C3&C4, F1&F2, F3&F4). We note that discriminatively trained CA-chain model through direct search

Table 5: Affordance classification performance trained on a large training set with 1400 images (700 with category labels) and 4200 affordance labels in total. Columns 2-7 are the testing error rate for individual, average, group affordances prediction and category recognition. The last two columns indicate the objective function being maximized and whether the category label is fixed in training (see also table 4). Error rates for CA approaches that are at least 3% better than the corresponding error in DP is highlighted. The smallest and the largest category error rates are also highlighted.

Training error (with large training set)								
Algorithm	Traverse	Move	Support	Average	Group	Category	LL/CLL	Fix C
DP	0.2%	4.1%	0.9%	1.7%	–	–	CLL	–
C1	4.3%	8.0%	6.8%	6.4%	14.7%	13.1%	LL	Yes
C2	3.8%	8.5%	6.6%	6.3%	14.9%	17.0%	LL	No
C3	1.2%	6.7%	2.2%	3.4%	9.2%	5.6%	CLL	Yes
C4	1.1%	6.9%	1.9%	3.3%	9.2%	6.3%	CLL	No
C5	0.4%	1.8%	1.3%	1.1%	3.1%	19.0%	CLL	No
F1	2.1%	2.2%	3.4%	2.6%	6.0%	12.9%	LL	Yes
F2	1.3%	2.0%	2.1%	1.8%	4.1%	15.7%	LL	No
F3	0.1%	0.8%	0.3%	0.4%	0.9%	1.9%	CLL	Yes
F4	0.1%	0.6%	0.3%	0.3%	0.7%	8.3%	CLL	No

Testing error (with large training set)								
Algorithm	Traverse	Move	Support	Average	Group	Category	LL/CLL	Fix C
DP	2.6%	4.0%	4.1%	3.6%	9.6%	–	CLL	–
C1	5.1%	7.5%	7.2%	6.6%	15.2%	14.6%	LL	Yes
C2	4.6%	8.0%	7.2%	6.6%	15.6%	18.6%	LL	No
C3	2.4%	6.2%	4.0%	4.2%	10.8%	9.4%	CLL	Yes
C4	2.3%	6.6%	3.6%	4.2%	10.8%	9.7%	CLL	No
C5	3.0%	2.4%	3.6%	3.0%	6.8%	17.7%	CLL	No
F1	3.4%	3.2%	5.0%	3.9%	9.3%	14.2%	LL	Yes
F2	3.1%	2.9%	4.7%	3.6%	8.1%	17.3%	LL	No
F3	2.1%	1.6%	2.9%	2.2%	5.2%	6.9%	CLL	Yes
F4	1.9%	2.0%	3.1%	2.3%	5.7%	11.8%	CLL	No

(C5) outperforms CA-chain models from other learning algorithms (C1-4) and even the generatively trained CA-full models (F1,F2). Although on the other hand, its category recognition error is among the highest: 19.0% in training and 17.7% in testing. This is because the category structures are altered only to favor affordance classification. For example, it is more beneficial to classify non-movable boxes as cabinets rather than chairs, because this "misclassification" of the object category indeed helps affordance classification. In fact, had the categories for all the testing data been classified correctly, the chain model can only classify both the non-movable chairs and boxes as movable, because for both

of these categories *majority* of the objects are movable. This fact is also observed by noting that while the category error is 17.7% in testing (C5), the group affordance error is significantly smaller as being 6.8%. This error rate suggests that all of the 3 affordances are predicted correctly for 93.2% of the testing images. Therefore it must be the case that some errors made in category recognition either do not result in mistakes about affordances (e.g. classify table as shelf) or even improve affordance classification (e.g. classify non-movable chair as cabinet).

The DP approach's classification performance is comparable to the best of the category models for individual affordance classification[1]. In fact, DP directly minimizes a loss function connected to the classification error for each individual affordance. The weakness of the DP approach is its ability to predict a group of affordances all correctly. The group affordance classification error is 9.6% in testing. This is greater than the discriminatively training CA-full models (F3,F4) and the discriminatively trained CA-chain model through direct optimization (C5). The major reason lies in the fact that multiple affordances are learned independently with DP, providing no direct procedure to minimize the group affordance classification error.

Finally, an interesting observation is from comparing the results of C4 and C5, both of which aimed at maximizing `CLL` for the CA-chain model. The only difference is that C4 maximizes `CLL` through EM while C5 maximizes it directly through searching. We observe that C4 is more likely to reach a local maximum close to the result of C3 where the category C labels are fixed in training — a local maximum that is more consistent with the true category-appearance model. On the other hand, the direct optimization approach in C5 without EM may reach a different local maximum with (usually) much higher category recognition errors but better accuracy on affordance prediction.

6.4 Learning with Small Training Set

In this experiment, we compare the performance of the learning algorithms on a small training set. It contains 420 images in total, 60 per category. Only 1 (out of 3) affordance

[1]We empirically discovered that 3 Gaussian components each for the positive and negative affordance class achieves the best classification accuracy on our data set.

Table 6: Affordance classification performance trained on a small training set with 420 images (210 with category labels) and 420 affordance labels in total. Only error rates that are at least 3% better than corresponding error in the DP approach is highlighted. The smallest and the largest category error rates are also highlighted.

Training error (with small training set)								
Algorithm	Traverse	Move	Support	Average	Group	Category	LL/CLL	Fix C
DP	0.0%	0.0%	0.0%	0.0%	–	–	CLL	–
C1	3.8%	11.8%	9.8%	8.1%	–	12.9%	LL	Yes
C2	5.0%	11.0%	9.0%	8.1%	–	16.2%	LL	No
C3	2.5%	7.9%	3.8%	4.5%	–	1.4%	CLL	Yes
C4	3.1%	7.9%	3.8%	4.8%	–	0.0%	CLL	No
C5	0.0%	0.8%	0.0%	0.2%	–	43.3%	CLL	No
F1	1.3%	1.6%	6.8%	3.1%	–	14.3%	LL	Yes
F2	0.0%	0.0%	0.8%	0.2%	–	16.2%	LL	No
F3	0.0%	0.0%	0.0%	0.0%	–	0.5%	CLL	Yes
F4	0.0%	0.0%	0.0%	0.0%	–	3.3%	CLL	No

Testing error (with small training set)								
Algorithm	Traverse	Move	Support	Average	Group	Category	LL/CLL	Fix C
DP	7.9%	9.2%	11.7%	9.6%	26.5%	–	CLL	–
C1	6.4%	8.8%	9.0%	8.1%	19.3%	17.2%	LL	Yes
C2	6.6%	9.1%	8.6%	8.1%	19.6%	16.8%	LL	No
C3	4.5%	9.0%	6.0%	6.5%	17.2%	12.1%	CLL	Yes
C4	4.3%	8.6%	6.4%	6.4%	17.4%	12.4%	CLL	No
C5	4.6%	7.9%	8.6%	7.0%	18.2%	43.5%	CLL	No
F1	5.6%	10.2%	10.3%	8.7%	23.1%	16.3%	LL	Yes
F2	4.2%	12.4%	10.7%	9.1%	24.8%	17.4%	LL	No
F3	3.9%	9.1%	8.6%	7.2%	19.9%	12.5%	CLL	Yes
F4	4.9%	9.1%	10.4%	8.1%	22.6%	16.9%	CLL	No

label is given for each image, therefore each affordance has approximately 140 training images. Category labels for half of the images (210) are provided for training.

Similarly, we observe that discriminative training in general outperforms generative training. Under the criteria of group affordance classification performance, almost all CA learning approaches significantly outperform the DP approach — we see this as one of the major advantage of the categorical approach over direct perception. This is because the errors occur (in theory) almost independently for each affordance in the DP approach, therefore the cumulative contribution from each affordance's error results in a high group affordance classification error. In fact, the estimated group affordance classification error ξ_∞ based on equation 66 is 26.2%, close to the true empirical error of 26.5%.

In training, 3 of the CA-full approaches (F2-4) achieve close to 0 training error. We attribute this largely to the discriminative power of the logistic regression classifiers embedded in the CA-full approach. This is especially true for the result of F2 with a high category error of 16.2%, because the classifiers provide a "second chance" to "correct" the category recognition errors that might possibly lead to mistakes in affordance classification. For example, even if a table instance is mistakenly classified as a chair, but within the chair category, it is classified as "supportive" chair, therefore achieves correct affordance prediction. However, the performance on testing set is different: the CA-chain models outperform both the CA-full approach and the DP approach (which also has 0 training error), implying over-fitting of the latter two approaches. In fact, the CA-full model is vulnerable to over-fitting in the EM framework, because at each iteration, even if the category is incorrectly classified, the within-category classifier is learned to correct this error in affordance prediction, therefore resulting in an incorrect posterior distribution of the categories ${q_{n,c}}^{(t)}$ — the belief about an incorrect category prediction is strengthened because the affordances are predicted correctly. Therefore it is importantly to initialize the training with a reasonable category model at step 3 in algorithms 1-3. We suggest to initialize with a generatively training CA-chain model from C1.

We also note from comparing discriminative training results (C3-4 & F3-4) that fixing the categories in training improves recognition performance over not-fixing. However, the difference is not large because of the following two reasons in implementation. First, even if training without fixing the categories, the initialization of the category-appearance model is always from the given category labels, which may be close to the local maximum that EM arrives. Second, the training algorithms perform random-start hill climbing in maximizing `CLL` with post model-selection step based on small affordance classification errors which may also coincide with small category error rate.

The conclusion for scarce data training is that discriminative training with the CA-chain approach (C3-5) outperforms both the CA-full approaches and the DP approach, both of which suffer from over-fitting the training data. Note the fact that the CA-chain model can be learned with good affordance classification performance implies the existence of an

categorization that is predictive for the affordances.

6.5 Further Discussion on CA Model Training

In this section, we closely examine several issues in CA model training, aiming at providing more insights from the experiments. Data in the following discussion are from section 6.4.

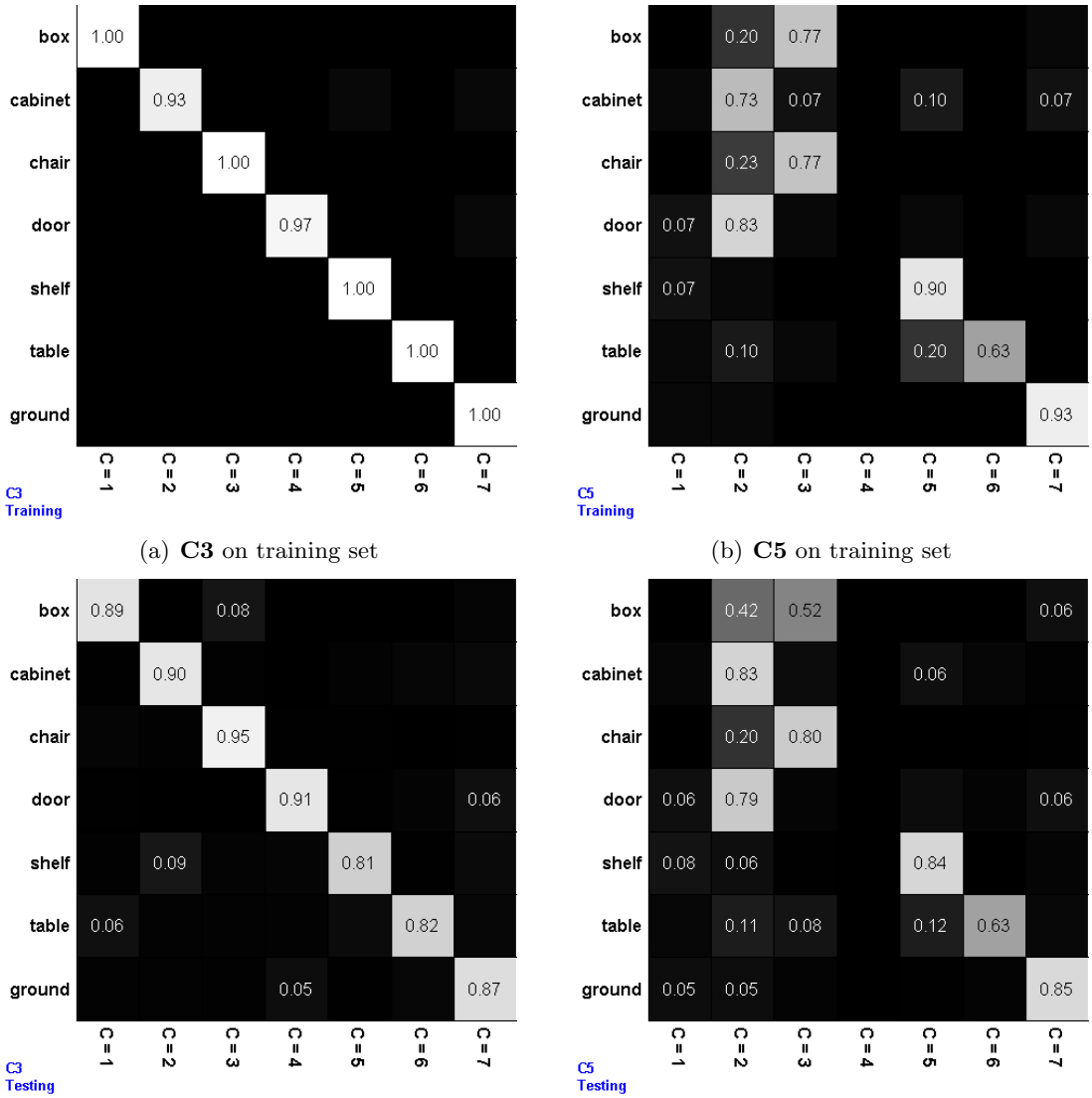
6.5.1 Generative vs. Discriminative Training

In appendix B, we show that for binary classifier learning in general, the negate of `CLL` is an upper-bound of the expected classification error. In this sense, maximizing `CLL` in discriminative training is effectively minimizing the error bound. On the other hand, we observe in generative training that although `LL` value increases at each EM iteration, `CLL` value often decreases — implying that the increase in `LL` is because of the increase in the likelihood of the appearance observation $P(X)$, which does not contribute to the classification performance. This intuitively explains our experimental observation that discriminative training outperforms generative training in terms of classification error. Detailed discussion with data illustration from the experiments are provided in appendix C for completeness.

6.5.2 The Process of Re-categorization

In table 6, it is interesting to note that the direct search based discriminative training algorithm C5 achieves very small training and testing error for affordances, but at the expense of a very high category recognition error of 43.3%. Compared to the classification result from C3, the best among all learning algorithms, C5 obtains comparable average affordance classification error of 7.0% (vs. 6.5%) and group classification error of 18.2% (vs. 17.2%), even if the category recognition error as high as 43.3% (vs. 12.1%).

To look closely into the difference of the two methods, we calculate the confusion matrix on category recognition for the models learned from both algorithms. In figure 18(a)(c), the confusion matrices from the category-affordance model learned with C3 is consistent with the "ground-truth" pre-defined categories. However, the confusion matrices from the model learned with C5 suggest high classification error as shown in figure 18(b)(d). We reorganize the confusion matrix to reflect the correspondence to the ground-truth categories in figure



(a) **C3** on training set

(b) **C5** on training set

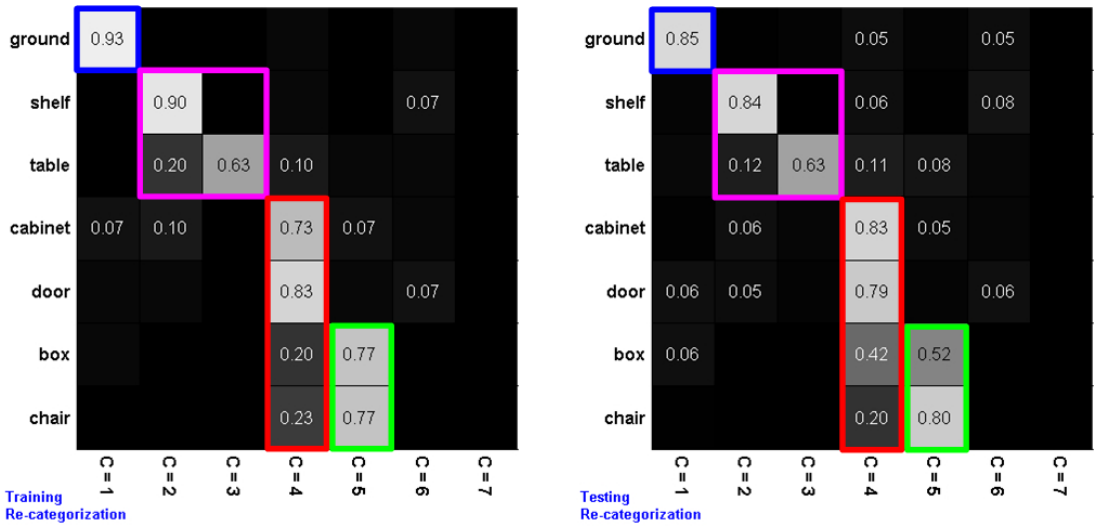
(c) **C3** on testing set

(d) **C5** on testing set

Figure 18: Confusion matrix on category recognition. (a)(c) Category model learned from C3 applied to training and testing sets respectively; (b)(d) Category model learned from C5 applied to training and testing sets. See also table 6.

19.

We see from figure 19 that the confusion indeed suggests a re-categorization of the original 7 object categories. Their affordance probabilities are shown in table 7. Compared with the pre-defined categorization, the categories defined after the re-categorization are summarized as follows: (1) non-movable boxes and chairs as well as doors and cabinets are now in the same category `(C=4)`, with the same non-traversable, non-movable and non-supportive affordances; (2) movable boxes and chairs construct a category `(C=5)` that is



(a) Recategorization by C5 on training set

(b) Recategorization by C5 on testing set

Figure 19: The confusion matrices illustrate the re-categorization implied by the CA-chain model learned with C5 approach. Objects are re-categorized in order to obtain correct affordance classification. Although the learned model still specifies 7 categories, category 6&7 are rare. The rectangles represents categories that have the same set of affordances values.

non-traversable, movable and non-supportive; (3) the boundaries between shelf and table are blurred reflected in two categories `(C=2,3)`, with 20% tables "misclassified" as shelf; (4) the ground is still a category by itself; (5) 3 categories `(C=4,6,7)` have the same set of affordance values.

We need to point out that it is because of the small training set size that the CA-chain model reaches a categorization where *category indeed determines affordance*, i.e. $P(A^k|C)$ is either 0 or 1. The exact learned $P(A^k|C)$ values in table 7 is different because for numerical stability in EM we had set a parameter ε to avoid the exact value of 0 or 1 in the model parameters. The learning rule for category-to-affordance probability parameter is:

$$\hat{P}(A^k = 1|C = c) = (\sum_{a_n^k=1} q_{n,c})/(\sum_{n \in S_k} q_{n,c}) \quad (67)$$

$$P(A^k = 1|C = c) = \max\Big(\varepsilon, \min(\hat{P}(A^k = 1|C), 1 - \varepsilon)\Big) \quad (68)$$

where $\hat{P}(\cdot)$ is the learned parameter for category-to-affordance probability and $P(\cdot)$ is the regularized model parameter. Recall that S_k is the training set for affordance k defined in equation 40, such that a_n^k for each training sample in S_k is observed to be either 0 or 1.

Table 7: Re-categorization learned from C5 algorithm. We list the category-to-affordance probability parameters learned in training and the correspondence between the new categories and the pre-defined categories. The new categories are constructed in such a way that they determine affordance values, which greatly facilitates affordance prediction.

Recategorization and prior P(C)		Learned $P(A^k = 1\|C)$ from C5			Corresponding Category	Posterior from data
		Traverse	Move	Support		
C = 1	19.5%	$1 - \varepsilon^*$	ε	ε	ground	15.2%
C = 2	13.9%	ε	ε	$1 - \varepsilon$	shelf/table	17.6%
C = 3	16.9%	ε	ε	$1 - \varepsilon$	table	9.1%
C = 4	10.5%	ε	ε	ε	cabinet/door/box/chair	30.1%
C = 5	11.5%	ε	$1 - \varepsilon$	ε	box/chair	23.8%
C = 6	11.2%	ε	ε	ε	random	3.3%
C = 7	16.5%	ε	ε	ε	none	0.0%
*: ε is defined to be 0.05 in training.						

This implies that the ground truth categorization — the one with the 7 categories we had defined — is not the *only* plausible categorization *under the task driven setting to predict the exact 3 affordances*. Empirically, the categorization learned from C5 is at least as good as the one learned from C3 in terms of the model's affordance prediction accuracy. Moreover, knowing the categories labels from the re-categorization is sufficient to determine the affordance values, which is not possible with the pre-defined categorization (at least for movable affordance). This is also observed in single affordance learning scenario that linguistic object categories may be irrelevant, as in the the slippery affordance classification work by by Angelova et. al. [4, 3], where they observed that learning the ground truth terrain categories of sand, soil, gravel, asphalt may not be necessary for learning the slippery affordance classifier. Trained generatively through the EM algorithm to maximize the joint likelihood of visual and mechanical inputs and the mechanical measurements (that is to be predicted), the maximum likelihood model indeed sacrifice the accuracy in terrain category recognition to better explain the mechanical behavior (i.e. slippery).

As we will discuss later in section 7, the new categorization does have higher category utility than the pre-defined one.

6.5.3 Interpreting the Model Parameters Learned Discriminatively

Another interesting observation from table 7 is made by comparing the model prior probability of the categories $P(C)$ (2nd column) and the posterior category frequency calculated

from training data (last column). The posterior category frequency shows the frequency of training data being classified as different categories (calculated based on the category posterior $p(c|x_n)$ for each sample). We see the prior and the posterior probabilities are significantly different. For example, the model prior for category 7 is 16.5% but none of the training data are classified as coming from this category. Category 6 also corresponds to a small proportion, i.e. 3.3% of the training data. The reason is that C5's direct optimization of the `CLL` function does not involve category learning; instead all the parameters can be tuned only to increase the `CLL` function. The learned prior of category 7 does not matter because its Gaussian mixtures are so far from the data distribution that almost no objects can be classified as from this category. This is less likely to occur for discriminative training with EM (C3), where the learning of category-appearance model at each iteration involves an intermediate step of learning a categorization to best match the posteriors.

6.6 New Affordance Learning

In this section, we compare the procedure and the performance of both the DP and the CA approach in learning a new affordance A^{K+1} on top of a prior model of the K pre-learned affordances. Because of the independent learning among the affordance models in the DP approach, there is little to benefit from a prior affordance knowledge-base. The DP approach thus requires as many training data as learning an affordance classifier from scratch.

However, in the CA-chain model, we can assume that the new affordance training data do not change the category-appearance model $P(X|C)$ — which is only related with category C and their appearance X and therefore is to be "shared" by the new affordance learning. In the CA-chain model, what remains to learn is the category-to-affordance probability $P(A^{K+1}|C)$. Learning such probability from a small training set is indeed possible.

6.6.1 New Affordances: Magnetic, Lockable and Flammable

To compare the learning algorithms in learning new affordances, we introduce three new affordances:

A^4. *Magnetic*: an object is considered magnetic if a magnetic sticker can be attached to it. All cabinets, shelves and elevator doors are magnetic.

A^5. *Lockable*: an object is lockable if it requires a key to access. Cabinets and office doors are lockable, but not elevator doors.

A^6. *Flammable*: an object is flammable if it can be ignited. Boxes, chairs, office doors, tables, and carpet type ground are flammable; while shelves, cabinets, elevator doors (all made of metal) and marble type ground are inflammable.

The three affordances are considered for learning new affordances with the robot, assuming they are relevant for an indoor robot to consider.[2]

Table 8: The set of new affordances and the pre-learned affordances. This is used to compare learning a new affordance with different training algorithms of direct perception and category-affordance models.

	Pre-learned affordance set			New affordance set		
Category	Traversable	Movable	Supportive	Magnetic	Lockable	Flammable
Box	No	Yes/No	No	No	No	Yes
Cabinet	No	No	No	Yes	Yes	No
Chair	No	Yes/No	No	No	No	Yes
Door	No	No	No	Yes/No	Yes/No	Yes/No
Shelf	No	No	Yes	Yes	No	No
Table	No	No	Yes	No	No	Yes
Ground	Yes	No	No	No	No	Yes/No

With a CA model already learned based on the previous set of affordances, learning the new affordances can be very efficient. We assume that the learning of the category-to-affordance inference is decoupled from learning the category-appearance model, the latter has already been learned from an existing set of affordances and remains the same. We demonstrate how this sharing of a categorical structure among affordances makes it possible to learn the new affordances efficiently.

[2]We have not yet implemented methods for the robot to automatically detect them but use manual labeling.

6.6.2 Learning with the CA-chain Model

We first consider learning the new affordances assuming that category sufficiently determines affordance, i.e. each $P(A^k|C)$ is close to 0 or 1. Theoretically if we had known this fact in training, only N training images are needed — one per category. In practise however, since whether this condition holds is unknown and that the category recognition performance is imperfect, we can estimate the $P(A^{K+1}|C)$ from the posterior of the categories $q_{n,c}$ of the training data as in equation 67.

We conduct experiments to learn each of the new affordances with both the DP approach and the CA-chain approach. For DP approach, we follow exactly same procedure because of the independence among learning different affordances. For the CA-chain approach, the category-affordance model is pre-learned from section 6.3 with a category recognition error rate of 6.9% (see table 5 F3).

Table 9: New affordance learning performance with changing training set size. The CA-chain approach is able to learn with very few training data, but the performance is limited with respect to whether categories sufficiently determines affordances.

Affordance	Model	Testing error (%) vs. size of training set				
		10	20	30	50	100
Magnetic	CA-chain	15.1 ± 8.9	5.8 ± 3.4	6.3 ± 5.6	4.6 ± 0.1	4.6 ± 0.0
	DP	—	28.9 ± 17.7	17.9 ± 4.6	12.9 ± 3.7	8.4 ± 1.5
Lockable	CA-chain	16.4 ± 8.6	6.7 ± 4.0	7.5 ± 4.5	5.6 ± 2.7	5.0 ± 0.1
	DP	—	19.8 ± 6.5	17.7 ± 4.1	12.4 ± 3.6	8.2 ± 1.8
Flammable	CA-chain	18.4 ± 7.2	10.4 ± 3.9	10.2 ± 5.0	8.6 ± 2.3	7.9 ± 0.1
	DP	—	28.4 ± 17.7	21.0 ± 6.7	14.4 ± 2.8	10.9 ± 2.4
Average of 3	CA-chain	16.7 ± 6.6	7.6 ± 2.3	8.0 ± 4.4	6.3 ± 1.2	5.8 ± 0.1
	DP	—	25.7 ± 11.9	18.8 ± 3.8	13.2 ± 2.0	9.2 ± 1.1

However, even if category recognition is perfect, the CA-chain approach is still bound to have affordance classification error for object categories that do not determine affordances. For example, the best prediction of the lockable affordance A^5 for a door object is lockable, assuming lockable doors are more frequent than unlockable doors. The classification error induced in this model is $P(\hat{C})P(A^5 = 0|\hat{C})$, where $\hat{C}$ denotes the door object. If the non-lockable ones in the door category is with relatively small probability $P(A^5 = 0|\hat{C})$, then the CA-chain model is still a good approximation. Otherwise, we need to break the sufficiency

assumption and learn a lockable classifier within the door category, the same technique we had used in the CA-full model. Even in this case, the learning approach in this section serves as a mechanism to *detect* when the sufficiency assumption breaks.

Table 9 list the testing error of classifying new affordance with model learned from training sets of different size. For a fixed training set size, 10 such training sets are drawn to compute the mean and standard deviation of the performance with the different learning approaches. We see that the CA-chain approach is able to learn from very few amount of training data, while the DP approach suffers significantly from a lack of training data. However, limited by the assumption that categories determines affordance which ignores the few exceptions of the objects in each category with different affordance values, the classification performance of the CA-chain approach is limited. In this case, the learned parameter value of $P(A^k|C)$ is used to detect when the assumption breaks and a category-specific affordance classifier can be learned.

6.6.3 Learning with the CA-full Model

As mentioned above, category-specific affordance classifiers can be learned when the sufficiency assumption breaks. This is the same learning procedure as in the CA-full learning algorithm, except that the category-appearance model learning is decoupled and assumed to be done already with existing set of affordances. Therefore we only need to estimate the category posterior $p(c|x_n)$ from existing model and learn the affordance classifiers with weighted training data per category, same way as described in equation 39 in section 5.2.

We compare this approach together with the CA-chain and the DP approaches, with the training set size ranges from 10 to 400. The experiment results are summarized in figure 20. As discussed, the CA-chain approach converges to its best performance very fast, but the error rate remains to be at a relatively high level — related to the degree to which the sufficiency assumption holds. The CA-full approach initially have a larger error rate because of over-fitting the training data with the additional DOF in the category-specific classifiers. With more training data, its performance improves and outperform the CA-chain model. In fact, although our experiment is using relatively simple logistic regression

classifiers, more complicated classifiers can be used — for example, the same one used in DP approach or a boosted ensemble. The DP approach can eventually achieve a smaller error than the CA-chain model and comparable to the CA-full approach as in figure 6, but its error rate with limited training data for new affordance learning is much higher than the CA-chain approach. This experiment suggests that utilizing a categorical representation of the objects make it possible to make reasonable generalization when the training data is scarce in learning a new affordance.

6.6.4 Discussion and Summary for New affordance Learning

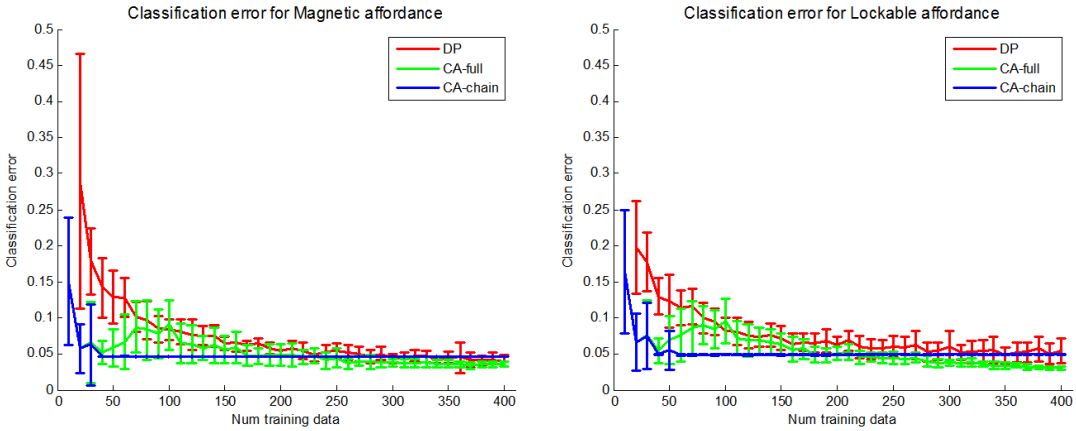
In this section we have demonstrate the learning new affordances with the categorical model. It can be summarized as a two-step procedure:

1. Learn with a CA-chain approximation of $P(A^{K+1}|C)$ for each category. If this value is not close to 0 or 1,

2. Learn a category-specific affordance classifier $P(A^{K+1}|C, X)$.

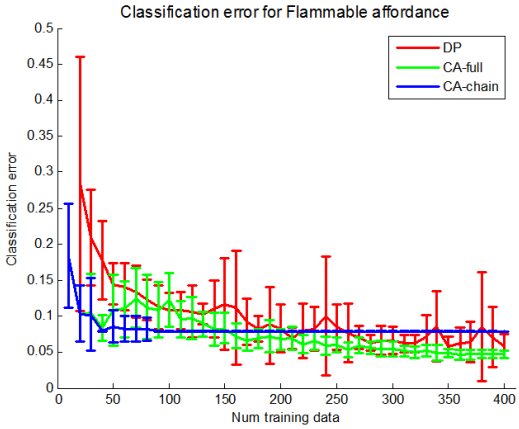
Intuitively the first step is to make assertions such as "chairs are all flammable" as long as the probability is very close to 1. The hope is that at least for some categories it is approximately correct to be able to generalize as such. For categories such that this approximation breaks, a classifier of flammable chairs vs non-flammable chairs are learned, but we only expect to do this for a few categories. In the case of very scarce training data, only the first step in training is needed.

A final comment is that so far we have only considered the case where for each training sample only the new affordance label is provided. If however, some other affordance labels are also provided in the training data of this new affordance, then the learning can be improved with the CA model, because more accurate category posterior $q_{n,c}$ can be estimated — not from appearance alone but also from other affordance labels. For example, this posterior is exactly $p(c|x_n)$ if no affordances other than the new one A^{K+1} is observed, but if pre-learned affordances are also observed, the posterior is now proportional to $p(c|x_n)p(a_n|x_n, c)$ therefore impacted by the accuracy of the classification on pre-learned affordances.

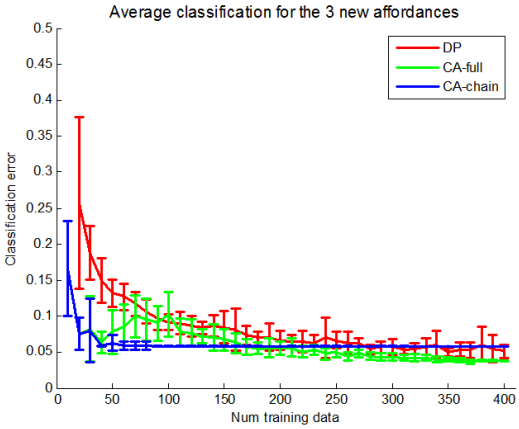


(a) Magnetic

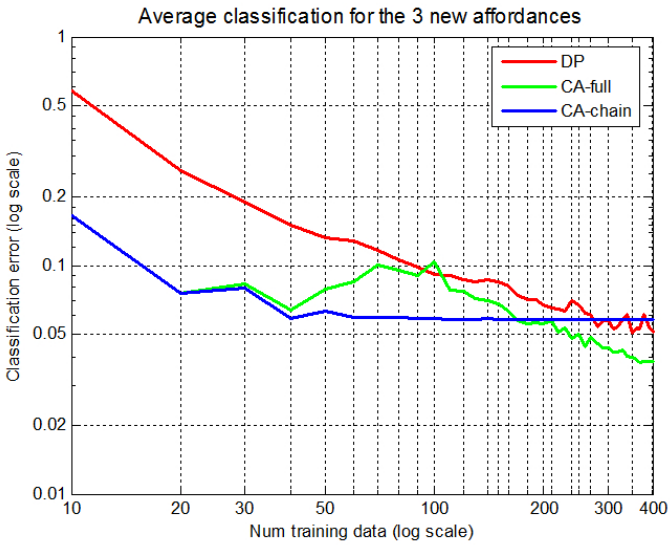
(b) Lockable



(c) Flammable



(d) Average of the three



(e) Average of the three (in log scale)

Figure 20: New affordance classification error on testing data, with models training with different size of training set. (a)(b)(c) Individual affordances; (d)(e) Average performance of the 3 affordances and its log-log plot.

6.7 Learning with Unbalanced Affordance Data

In practise, there could be cases when some affordances labels are more difficult to obtain than others. For example, while traversability is can be obtained as the robot drives, the supportive affordance requires the robot to conduct more complicated experiments. On other occasions, the affordance to learn could exactly be the situations that the robot attempts to avoid (such as being hazard). Therefore the size of the training set for affordances are unbalanced; some affordance training data may be rare. Techniques for learning rare event detectors, such as the cascade classifier, normally require a large amount of training data [92, 99]. The CA model learns these affordances by using other affordances to learn the categorical model and by treating the rare affordance as a new affordance with scarce training data. In other words, since this particular affordance labels are rare, its impact on category-appearance learning can be neglected, therefore the whole learning process is equivalent as learning a CA model with other affordances and then learn this rare one as a new affordance. Experimental demonstrations of this idea is similar to that of the new affordance learning section and omitted to avoid redundancy.

6.8 Further Discussion and Conclusions

In this section, we discuss the conclusions drawn from the experiments, with insights about applying affordance learning in building a robot platform.

The power of direct perception For single affordance learning, the direct perception approach, although not making use of the categories of objects, is a straightforward approach with reasonable performance. We have demonstrate evidence in this chapter with the experiment of affordance learning from a large training set as well as in chapter 3 with the tasks of learning traversability and preferability. The advantage of direct perception is its directness, which assumes no specific structure of the world (e.g. categorical) but targets directly to the learning goal of minimizing the affordance prediction error. In the scenario learning a single affordance with few domain knowledge but abundant training data, we recommend that a direct perception approach to be adopted.

The use of categorization We've shown that by making use of the categorization of objects, affordance learning can be decoupled into learning this categorization and the affordance classifiers for each category. In the case when the categorization is reliably learned, affordance learning can be made more efficient. The fact is that some object categories almost surely determines some affordances, such as the table being non-traversable. Observing the affordance of an object *instance* enables the learner to make assertions about the affordance properties of the *category* to which this instance belongs, therefore making "a-few-shot" learning possible. When this is not the case, for example with the movable affordance for the chair category, learning chairs that are movable against non-movable can be expected to be easier than learning the distinction between all movable and non-movable objects in general. *Category-specific* features or decision rules may be found to predict certain affordances, which do not need to hold *universally* for all objects. In the chair example, the existence of wheels is a key feature for predicting the movability of chairs, which does not, and also does not need to, apply universally to all objects. The limited scale learning task of category-specific affordance classifiers rather than a universal classifier can be regarded as a divide-and-conquer approach. In this chapter, we demonstrate the advantage of the category-affordance model with experiments of affordance learning with scarce training data, new affordance learning, and group affordance prediction.

The hybrid category-affordance model The category-affordance model is a hybrid model in that the categorization is modeled generatively (e.g. with Gaussian mixtures) and the appearance-to-affordance classifiers is modeled discriminatively (e.g. with logistic regression). In this way, standard object category learning can be applied and object categories can be added (into the model) incrementally. The appearance-to-affordance classifiers are modeled discriminatively so that it is the object categories rather than the affordances that determine appearance — a constraint in the model design so as to "shared" the categorization model among different affordances.

The applicability of discriminative training For the category-affordance models, the discriminative training of the affordances leads to (also) the discriminative training of the

object categorization and the same holds for generative training. In the experiments, we observe that discriminative training in general outperforms generative training in terms of affordance classification error. Theoretically, this can be attribute to the fact that `CLL` function is an upper-bound of the expected classification error. Therefore, we conclude that in practice discriminative training should be adopted. On the other hand, the drawback is that discriminatively learned model parameters do not describe the data according to the actual distribution, but rather serves solely for the goal of the best classification error. In this sense, these parameters may be counter-intuitive (see section 6.5.3). Further more, if new object categories are expected to be added, discriminative training may not be applicable as it requires all existing categories be re-learned.

The CA-chain model and categorization In section 6.5.2, we show that learning the CA-chain model for affordance prediction is equivalent to categorizing the objects according to the affordances. The goal of minimizing affordance prediction error naturally results in the process of organizing objects into different categories with each of them being more predictive about the affordances. In the experiments, we see that it is plausible to organize non-movable chairs in the same category with cabinets, because they are consistent in predicting the first 3 affordances that we consider. This may not be optimal with a expanded set of 6 affordances as we later show in section 7.1.4. One could say that the particular categorization learned from C5 is "good" for predicting the first 3 affordances, but not good for the set of 6 affordances. How to assess the "goodness" of categorizations is the major topic of the next chapter, in which we discuss about the category utility in the affordance learning task setting.

CHAPTER VII

CATEGORY UTILITY AND AFFORDANCE PREDICTION

The category utility function in [27, 20] is defined to measure the "goodness" of a categorization, considering the predictive power of the categories to the attributes. In the task driven setting of affordance prediction, we limit the attributes to the affordances being considered, based on the fact that appearances are always observable and hence unnecessary to predict. This enable us to assess and compare the goodness of different categorizations, which reflects and also helps to better understand some claims and conjectures in the cognitive science research [10, 54, 30]. This also provides a different perspective on how to determine what categories are important to learn.

In this chapter, we first show that category utility function from [27, 20] can be divided into two parts: the model utility that is categorization dependent and the base utility that is not. The model utility are shown to be directly connected to affordance prediction with the CA-chain model, serving as an upper-bound of affordance prediction error. We then introduce the empirical utility that measures empirically the goodness of a categorization *model* being applied to a particular set of *data*. We compare the category utility of the CA models learned (with the 3 affordances) in chapter 6 on both the training data and the testing data, and furthermore the new affordance data set. This illustrates that an aggressive re-categorization (notably in the C5 approach) may over-fit the training data which in turn results in poor generalization ability for new affordances.

Finally, we further extend the utility measure by incorporating the fact that category recognition are not error-free. We re-derive the category utility in its matrix notation and formulate the empirical utility with the category recognition *confusion matrix*. This enables us to reconsider the question as to what object categorization are "optimal" to train a robot agent for affordance learning.

7.1 Analysis and Extension of the Category Utility Function

In this section, we first introduce the category utility function with respect to affordances and explain the implication of each term in the function. Then we connect the category utility function to the affordance prediction performance of the CA model, and make the distinction between the model utility and the empirical utility. We demonstrate these concept with data from previously discussed experiments in chapter 6.

7.1.1 The Gluck Category Utility

Consider a categorization (i.e. a partition) $C = \{C_n\}, n = 1, 2, ..N$ and a set of affordances $A = \{A^k\}, k = 1, 2, ...K$, where each affordance takes binary values $A^k = \{0, 1\}$. The Gluck category utility is:

$$\begin{aligned} \mathcal{U}_G(C; A) &\triangleq \sum_{n=1}^{N} p(C_n) \sum_{k=1}^{K} \left[\sum_{a=\{0,1\}} p(A^k = a|C_n)^2 - p(A^k = a)^2 \right] \\ &= \sum_{k=1}^{K} \left\{ \left[\sum_{n=1}^{N} p(C_n) \sum_{a=\{0,1\}} p(A^k = a|C_n)^2 \right] - \sum_{a=\{0,1\}} p(A^k = a)^2 \right\} \\ &\equiv \sum_{k=1}^{K} \mathcal{U}_G(C; A^k) \end{aligned} \tag{69}$$

where $\mathcal{U}_G(C; A^k)$ is defined to be the term in the parentheses $\{\cdot\}$, which corresponds to the category utility considering one affordance A^k. It is also known as the Goodman-Kruskal association index [28, 57]:

$$\mathcal{U}_G(C; A^k) \triangleq \left[\sum_{n=1}^{N} p(C_n) \sum_{a=\{0,1\}} p(A^k = a|C_n)^2 \right] - \sum_{a=\{0,1\}} p(A^k = a)^2 \tag{70}$$

This category utility definition always favors fine scale of categorization. In an extreme case, given a set of data, the utility is maximized when each sample point is itself a category. Therefore, in [27] the utility is multiplied by a factor of $\frac{1}{N}$ as a rough attempt to avoid over-fitting or the unlimited increase of number of categories N. For the ease of our discussion we refer to the value without the normalization factor as the Gluck utility, while the value with the normalization as the average Gluck utility.

7.1.2 Base Utility, Model Utility and Classification Error

We define the two terms in the Gluck utility definition in equation 70 as follows:

$$\mathcal{U}_M(C;A^k) \triangleq \sum_{n=1}^{N} p(C_n) \sum_{a=\{0,1\}} p(A^k=a|C_n)^2 \tag{71}$$

$$\mathcal{U}_0(A^k) \triangleq \sum_{a=\{0,1\}} p(A^k=a)^2 \tag{72}$$

Similarly, the utility $\mathcal{U}_0(A)$ and $\mathcal{U}_M(C;A)$ are the sum of $\mathcal{U}_0(A^k)$ and $\mathcal{U}_M(C;A^k)$ over k respectively. In [20], it has been discussed that $\mathcal{U}_0(A^k)$ is the error rate of affordance under *probabilistic prediction*. For example, if $p(A^k=1)=0.6$, we predict at 60% of the time at random that the affordance is true, and at 40% of the time that the affordance is false. Theoretically, the correct prediction rate is $0.6^2+0.4^2=52\%$. Similarly, $\mathcal{U}_M(C;A^k)$ is calculated with same probabilistic prediction within each category, such that the prediction is based on the probability $p(A^k=1|C_n)$ for each category, which is then summarized across all categories weighted by its priors $p(C_n)$. It can be summarized that $\mathcal{U}_0(A^k)$ is the *predictiveness* of the affordance distribution itself and $\mathcal{U}_M(C;A^k)$ is the predictiveness of the affordance based on a categorization.

Base utility However, what has not been explicitly discussed before is that $\mathcal{U}_0(A^k)$ also implies (a family of) categorizations that is independent from A^k. In that case each term $p(A^k=1|C_n)$ in $\mathcal{U}_M(C;A^k)$ is exactly $p(A^k=1)$, resulting in the form of $\mathcal{U}_0(A^k)$. It can also be considered as a special categorization with only one category. We refer to this quantity as the *base utility*. Because the Gluck utility is always non-negative, we have:

Claim 1 *The* base utility *measures the* minimum *predictiveness of the affordances based on a categorization, in which case it is independent from the affordances.*

Model utility Since the base utility is a constant that does not depend on specific categorization, in evaluating the "goodness" of a particular categorization, we are more interested in the $\mathcal{U}_M(C;A^k)$ term. $\mathcal{U}_M(C;A^k)$ is determined by the prior category distribution $p(C)$ and the category-to-affordance distribution $p(A^k|C)$, which, together with the category appearance model $p(X|C)$ specifies a CA-chain model previously discussed. In other words,

given a CA-chain model, we can evaluate the category utility of the categorization implied in the model — hence we refer to it as the *model utility*. The model utility is directly connected to the affordance prediction performance of the CA-chain model. Assuming that the model makes *no category recognition error* at all, we show that the affordance classification error $\mathbf{E}[\varepsilon_k]$ for each individual affordance A^k is bounded by the model utility defined on that affordance.

Claim 2 *The negative* model utility *(plus 1) is an upper-bound of the affordance classification error assuming no category recognition error.*

Proof Defining $p_{k,n} \triangleq p(A^k = 1|C_n)$, therefore $1 - p_{k,n} = p(A^k = 0|C_n)$. We note that the affordance prediction error of the CA-model within each category is the smaller value of $p_{k,n}$ and $1 - p_{k,n}$. We have:

$$\begin{aligned}
\mathbf{E}[\varepsilon_k] &= \sum_{n=1}^{N} p(C_n) \min(p_{k,n}, 1 - p_{k,n}) \\
&\leq \sum_{n=1}^{N} p(C_n) \min(p_{k,n}, 1 - p_{k,n}) [2 \max(p_{k,n}, 1 - p_{k,n})] \\
&= \sum_{n=1}^{N} p(C_n) 2 p_{k,n} (1 - p_{k,n}) \\
&= \sum_{n=1}^{N} p(C_n) \left[(p_{k,n} + 1 - p_{k,n})^2 - p_{k,n}^2 - (1 - p_{k,n})^2 \right] \\
&= 1 - \sum_{n=1}^{N} p(C_n) \left[p_{k,n}^2 + (1 - p_{k,n})^2 \right] \\
&= 1 - \mathcal{U}_M(C; A^k)
\end{aligned} \tag{73}$$

The second step holds because $\max(p_{k,n}, 1 - p_{k,n}) \geq \frac{1}{2}$. ∎

This can be similarly extended to multiple affordances, in which case we average the category utility and affordance classification over the number of affordances, resulting in:

$$\begin{aligned}
\mathbf{E}[\varepsilon] &\equiv \frac{1}{K} \sum_{k} \mathbf{E}[\varepsilon_k] \\
&\leq 1 - \frac{1}{K} \mathcal{U}_M(C; A)
\end{aligned} \tag{74}$$

The result explains the experiment in section 6.5.2, that discriminative training reaches a re-categorization with high category utility in training, because both the `CLL` and the

category utility are upper-bounds of the affordance classification error. Back to the Gluck utility, we have:

Claim 3 *The* Gluck utility *of a categorization w.r.t a set of affordances equals the difference between* model utility *and* base utility.

7.1.3 Empirical Utility

Previously discussed section 6.5.3, the probabilities specified by the CA model may not equal the posterior probabilities implied by the data. As we showed for example, this can be the case when the model is learned discriminatively. Consider a categorization model $P(C, X)$ and a (large) set of weakly labeled data $\{(x_n, a_n)\}$. Suppose that $\hat{p}(C)$ and $\hat{p}(A^k|C)$ are the probabilities estimated from the posterior density $\hat{p}(c|x_n)$ of each data, we can calculate the category utility from the real data distribution. Since $\hat{p}(A^k|C)$ are calculated from the data, this is the same process as learning a new affordance with the CA-chain model, because we are interested in the utility of the categorization which is sufficiently determined by $p(C, X)$.

The base utility remains the same assuming the marginal distribution of affordances (estimated from the large data set) do match with the real distribution. On the other hand, the category posterior on the test data effective implies a potentially different categorization from the model description (e.g. table 7 in chapter 6). Therefore the model utility is now replaced by what we refer to as the *empirical utility*, which is calculated in the same way as equation 71, but only from on the posterior probabilities. Similar as in section 7.1.2, the (negative) empirical utility is directly connected to the classification error as a upper-bound. An example of this is shown in figure 22, discussed in detail in the following section.

7.1.4 A Comparison of CA Models on Category Utility

We calculate the category utility for the categorizations learned in the CA-models discussed in sections 6.3 and 6.4. The purpose is to evaluate the "goodness" of these different categorizations in different settings, which helps understanding as what categorizations are more predictive than others in the task of affordance prediction.

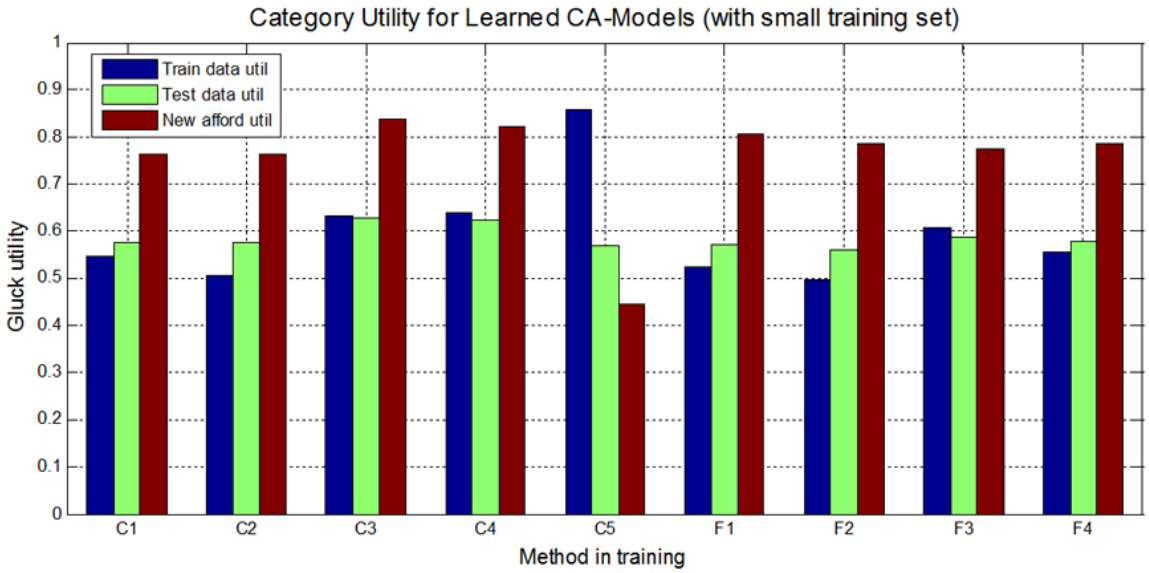


Figure 21: Category utility measured on the CA-models learned in previous experiment of training with small data set. We measure the Gluck utility with the training and the testing data of the initial set of three affordances (traversable, movable, supportive) as well as the utility with the new affordance set (lockable, magnetic, flammable).

Utility on initial affordance set We calculate the Gluck utility of the 9 CA-chain models learned in section 6.4, but with the model utility replaced by the empirical utility. For each model, we calculate the utility with the training and the testing data of the initial set of three affordances (traversable, movable, supportive) as well as the utility with the new affordance set (lockable, magnetic, flammable). For the new affordance case, we learn the category-to-affordance distribution with the existing categorization on a large training set and calculate the Gluck utility on the same set. This measures the intrinsic utility of the CA models, instead of affected by the possibly biased distribution with a small set of training data. All the utility calculations are based on the posterior distribution of $p(C|X)$ — on top of which $p(C)$ and $p(A^k|C)$ are calculated with the affordance labels being provided. Calculated on the posterior of the data set, this utility measure directly connect to the classification error, unlike the model utility.

The results are summarized in figure 21. The Gluck utility is calculated as the difference between empirical utility and base utility, measuring the increased predictiveness of affordances from a categorization. We first observe that the categorizations learned from

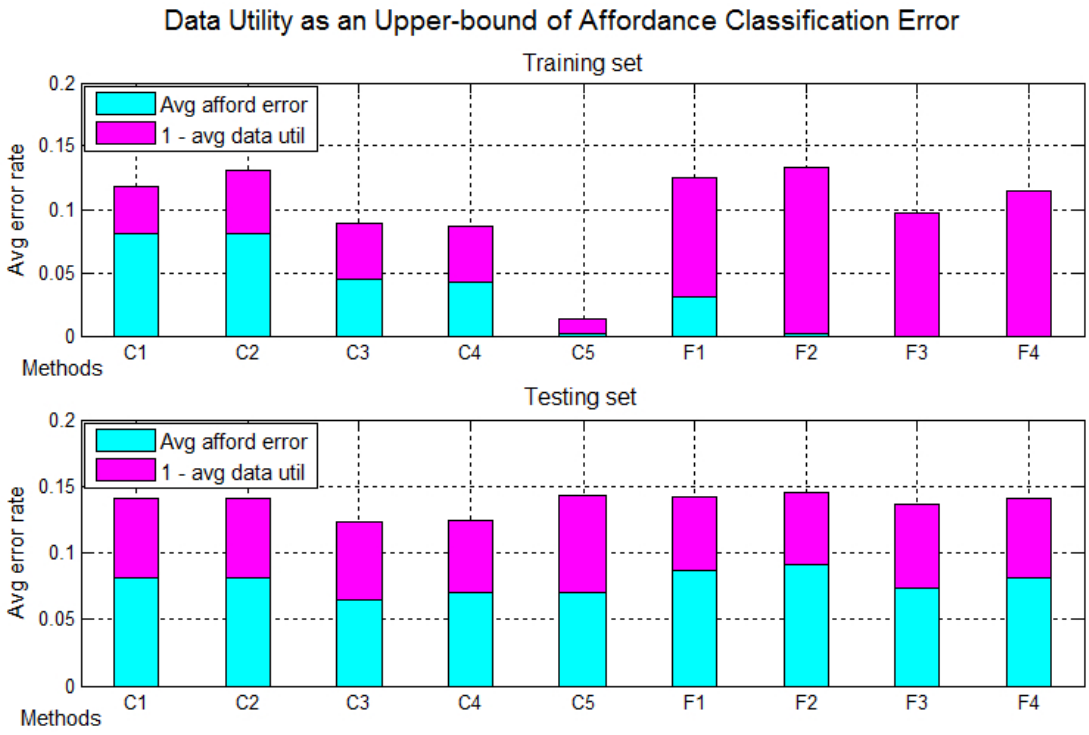


Figure 22: Empirical utility as an upper-bound of affordance classification error. For CA-full models, the classification error is much smaller because of the discriminative classifiers trained within teach categories, which goes beyond purely category-to-affordance prediction. The direct optimization approach C5 achieves very high category utility (and small affordance error) in training but not in testing, implying over-fitting.

the CA-full approaches have smaller utilities than those from the chain approaches (F1-4 vs. C1-4). This is because that unlike the CA-chain approach whose affordance classification error is directly related to the category utility (as an error-bound), the CA-full model's affordance classification performance is further greatly improved by the category specific classifiers. As a result, the utility measure is not as important for the CA-full approaches than the CA-chain approaches. Therefore, we will concentrate on discussion with the CA-chain models.

Second, we notice that on both the training and the testing sets, the generatively trained categorizations in general have smaller utility than their discriminative counterparts (C1-2 vs. C3-4). The re-categorization results from the C5 approach (direct optimization of `CLL`) has the highest utility on training data — much higher than other approaches — directly corresponding to its minimum training error among all the CA-chain models (see table 6).

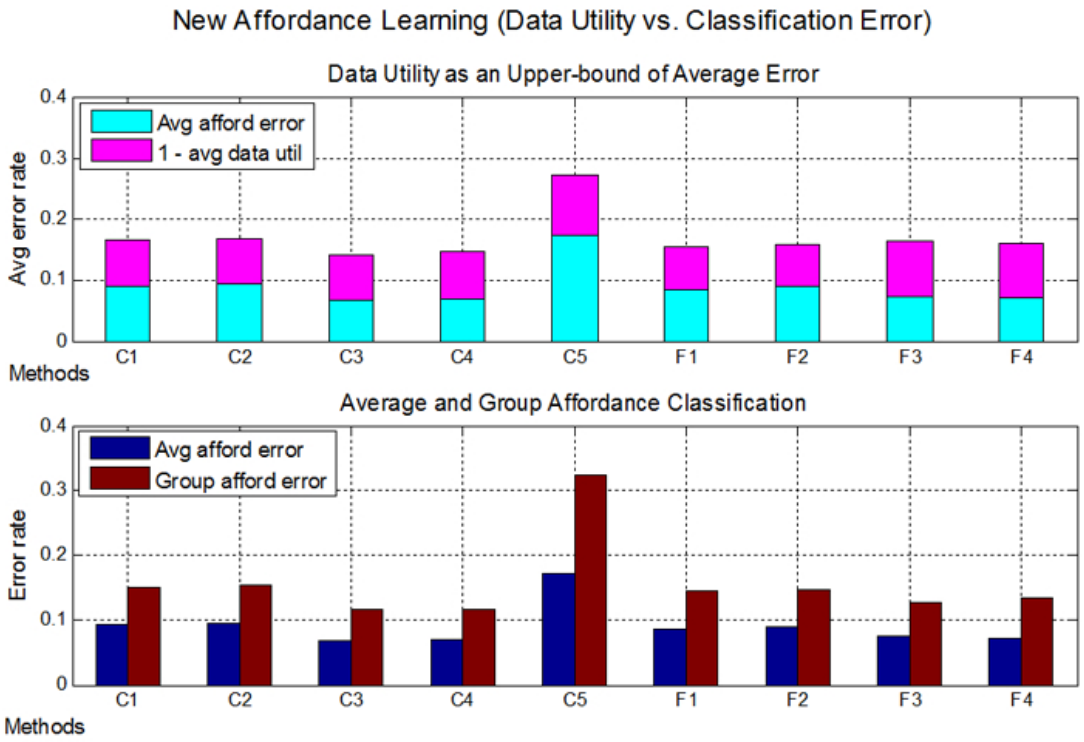


Figure 23: The CA models learned previously are applied to new affordance learning. The performance of C5 is much worse compared to other methods, implying that its re-categorization is not as suitable for the new affordance set. Top figure shows the empirical utility and the average classification error comparison on the new affordance set. Bottom figure shows both the average affordance classification error and the group affordance classification error.

In terms of the testing data, the categorization learned in C5 does not have the highest utility. This agrees with our observation on the affordance testing error presented in table 6, that C5 does not have the best testing error, which implies over-fitting.

We also show the figure of empirical utility and the affordance classification error (both averaged over 3 affordances) in figure 22. It can be seen that the empirical utility is indeed an upper-bound of the affordance classification error. As aforementioned, the affordance classification error with the CA-full models are very small not because of the category utility alone, but also because of the within-category classifiers. The over-fitting of C5 is also apparent in figure 22.

Utility on new affordance set What is more interesting is to compare the category utility with the new set of affordances, but not the affordance set that these CA-models are

trained on. While other approaches (C1-4&F1-4) have comparable category utility (as they also have comparable category recognition error), C5 has a very low category utility with respect to the new affordance set (see figure 21). This suggests while the categorization learned from C5 might be optimal (among other approaches) for the initial affordance set, it does not have good generalization for the new affordances. Because optimization in C5 has aggressively re-categorize the objects only to improve affordance classification on the initial affordance set, it is suboptimal for a different affordance set. For example, C5 had put non-movable boxes and chairs together with cabinets and doors in the same category (figure 19), which is beneficial for traversable/movable/supportive classification, but is definitely suboptimal for lockable (true for some doors) or magnetic (true for cabinets and some doors) affordances. In fact, it can be expected that had C5 been initially applied on a training set with 6 affordances altogether, it will probably not arrive at the re-categorization in figure 19.

Table 10: Model utility comparison for different object categorization. The ground truth model parameters $P(C)$ and $P(A^k|C)$ are listed together with the computed model utility. The 7-category-categorization has higher model utility than the 2-category-categorization. A^1 through A^6 are the 6 affordances listed in table 8.

Categorization with 7 categories					$\mathcal{U}_M(C;A) = 5.75$		
Category	$P(C)$	$P(A^k = 1\|C)$					
		A^1	A^2	A^3	A^4	A^5	A^6
Box	0.143	0	0.717	0	0	0	1.00
Cabinet	0.143	0	0	0	1.00	1.00	0
Chair	0.143	0	0.841	0	0	0	1.00
Door	0.143	0	0	0	0.126	0.874	0.874
Shelf	0.143	0	0	1.00	1.00	0	0
Table	0.143	0	0	1.00	0	0	1.00
Ground	0.143	1.00	0	0	0	0	0.740

Categorization with 2 categories					$\mathcal{U}_M(C;A) = 4.09$		
Category	$P(C)$	$P(A^k = 1\|C)$					
		A^1	A^2	A^3	A^4	A^5	A^6
Furniture	0.571	0	0.210	0.500	0.500	0.250	0.500
Non-furniture	0.429	0.333	0.239	0	0.042	0.292	0.872

Model utility with different categorization As an illustrative example, we compare the model utility with two different categorization: one categorization consists of the 7 object categories that we previously specified, the other categorization consists of two object categories – furniture (including cabinet, chair, shelf, table) and non-furniture (corresponding to box, door, ground). Recall that the model utility measures the predictive power of the categorization for the affordances in the ideal case with no category recognition error. In table 10 we list the ground truth model parameters of $P(C)$ and $P(A^k|C)$, from which the model utility are computed using equation 71. As we can see, the categorization of furniture and non-furniture does not have sufficient predictive power as compared with the 7 object categories – the model utilities are 4.09 versus 5.75 for the 2 categorizations respectively. For detailed discussions on utility comparison with different object categorizations, the reader may refer to [27, 20].

7.2 More on Category Utility

In previous discussion, we have enumerated 4 different utility definition: the Gluck utility, base utility, model utility and empirical utility. The distinction is made to connect category utility to the task of learning object categories for affordance prediction. We also provided experimental results on the initial affordance set and the new affordance set, which suggest that the "goodness" of an object categorization is *well defined* with respect to a set of affordances. Categorizations with higher utility on one set of affordances does not guarantee higher utility also on a different affordance set. Therefore, in training an agent with category based affordance classification, we have to both consider the current target affordance set as well as future affordances that the agent *may* need to learn. Because of this consideration for possible future affordances currently unknown, it is difficult to assert that a particular categorization is optimal. However, we could hypothesize that natural object categories defined by human is a reasonable choice.

In this section, we also provide a new derivation for these utility definitions in matrix form, which allows us to later discuss its implication in training an agent for the task of category based affordance classification.

7.2.1 Category Utility in Matrix Form

To facilitate further discussion of the category utility and empirical affordance prediction error on a particular data distribution, we derivation a new characterization of the category utility in matrix form. A number of matrix are defined as follows:

$$\mathbf{P} \triangleq [p(C_1) \quad ... \quad p(C_N)]^T \in \mathbf{R^N} \tag{75}$$

$$\mathbf{S_k} \triangleq \left[p(A^k = 1|C_1) \quad ... \quad p(A^k = 1|C_N)\right]^T \in \mathbf{R^N} \tag{76}$$

$$\mathbf{S} \triangleq [\mathbf{S_1} \quad ... \quad \mathbf{S_K}] \in \mathbf{R^{N \times K}} \tag{77}$$

also let $\mathtt{diag}(\mathbf{P})$ denote the square matrix with elements of $\mathbf{P}$ in the diagonal. With some matrix calculation we see that the Gluck category utility $\mathcal{U}_G(C; A^k)$ defined on one affordance in equation 70 can be equivalently written in the matrix form of:

$$\mathcal{U}_M(C; A^k) \equiv \mathbf{S_k}^T \mathtt{diag}(\mathbf{P}) \mathbf{S_k} + (\mathbf{1} - \mathbf{S_k})^T \mathtt{diag}(\mathbf{P})(\mathbf{1} - \mathbf{S_k}) \tag{78}$$

$$\mathcal{U}_0(A^k) \equiv \mathbf{S_k}^T \mathbf{P}\mathbf{P}^T \mathbf{S_k} + (\mathbf{1} - \mathbf{S_k})^T \mathbf{P}\mathbf{P}^T (\mathbf{1} - \mathbf{S_k}) \tag{79}$$

$$\begin{aligned} \mathcal{U}_G(C; A^k) \equiv\ & \mathbf{S_k}^T \left(\mathtt{diag}(\mathbf{P}) - \mathbf{P}\mathbf{P}^T\right) \mathbf{S_k} \\ & + (\mathbf{1} - \mathbf{S_k})^T \left(\mathtt{diag}(\mathbf{P}) - \mathbf{P}\mathbf{P}^T\right) (\mathbf{1} - \mathbf{S_k}) \end{aligned} \tag{80}$$

Note that both $\mathtt{diag}(\mathbf{P})$ and $\mathbf{P}\mathbf{P}^T$ are semi positive definite (SPD) matrices. In appendix E, we show that the difference of $\mathtt{diag}(\mathbf{P}) - \mathbf{P}\mathbf{P}^T$ is in fact also a SPD matrix, therefore the Gluck category utility is always positive.

The two terms we defined in 80 directly correspond to the two terms from the definition of Goodman-Kruskal association index in equation 70. It is not difficult to see because the probability $p(A^k)$ is specified given $\mathbf{S_k}$ and $\mathbf{P}$:

$$\begin{cases} p(A^k = 1) = \mathbf{S_k}^T \mathbf{P} \\ p(A^k = 0) = (\mathbf{1} - \mathbf{S_k})^T \mathbf{P} \end{cases} \tag{81}$$

Similar as in single affordance case, we can show that the model utility, base utility and the Gluck utility on the complete affordance set are:

$$\mathcal{U}_M(C; A) \equiv \mathtt{tr}\left[\mathbf{S}^T \mathtt{diag}(\mathbf{P}) \mathbf{S} + (\mathbf{1} - \mathbf{S})^T \mathtt{diag}(\mathbf{P})(\mathbf{1} - \mathbf{S})\right] \tag{82}$$

$$\mathcal{U}_0(A) \equiv \mathtt{tr}\left[\mathbf{S}^T\mathbf{P}\mathbf{P}^T\mathbf{S} + (\mathbf{1}-\mathbf{S})^T\mathbf{P}\mathbf{P}^T(\mathbf{1}-\mathbf{S})\right] \tag{83}$$

$$\begin{aligned}\mathcal{U}_G(C;A) \equiv \mathtt{tr}\Big[&\mathbf{S}^T\left(\mathtt{diag}(\mathbf{P}) - \mathbf{P}\mathbf{P}^T\right)\mathbf{S} \\ &+(\mathbf{1}-\mathbf{S})^T\left(\mathtt{diag}(\mathbf{P}) - \mathbf{P}\mathbf{P}^T\right)(\mathbf{1}-\mathbf{S})\Big]\end{aligned} \tag{84}$$

7.2.2 Empirical Utility and Category Recognition Confusion Matrix

We've defined empirical utility in section 7.1.3, which is calculated by applying the CA model on a particular data set. Defining $\hat{\mathbf{P}}$ and $\hat{\mathbf{S}}$ similarly as equations 75, 76, and 77 but on the posterior distribution calculated by the model, the empirical utility can be also expressed in matrix form:

$$\mathcal{U}_D(C;A) \triangleq \mathtt{tr}\left(\hat{\mathbf{S}}^T\mathtt{diag}(\hat{\mathbf{P}})\hat{\mathbf{S}} + (\mathbf{1}-\hat{\mathbf{S}})^T\mathtt{diag}(\hat{\mathbf{P}})(\mathbf{1}-\hat{\mathbf{S}})\right) \tag{85}$$

We then connect this empirical utility with the training process of a CA model. Suppose that the learner is provided with training examples with category labels from an optimal "ground truth" categorization. This categorization is chosen such that it has the highest category utility among all different categorizations with the same number of categories N. As aforementioned this is the model utility assuming no category recognition error, which is also the highest utility that learner could possibly achieve in supervised learning of the CA-chain model.

However, due to category recognition errors, the learned categorization does not achieve the "optimal" model utility. Given the confusion matrix of category recognition, we can approximately calculate the empirical utility. Let $\mathbf{F}$ be the confusion matrix, such that $\mathbf{F}_{i,j} = p(\hat{C}_j|C_i)$, with C_i being the correct category and $\hat{C}_j$ being the (mis)classification. Note that each row of $\mathbf{F}$ sum to 1, and each column $\mathbf{F_j}$ is the probability of all the ground-truth category being "mis-classified" as one particular new category $\hat{C}_j$. Leaving the proof in appendix F, we state the following claim about empirical utility:

Claim 4 *The empirical utility of a categorization in the existence of category recognition error (with confusion matrix* $\mathbf{F}$*) is:*

$$\begin{aligned}\mathcal{U}_D(C;A) &\equiv \mathit{tr}\left(\mathbf{S}^T\mathbf{B}\mathbf{S} + (\mathbf{1}-\mathbf{S})^T\mathbf{B}(\mathbf{1}-\mathbf{S})\right) \\ \mathit{where}\ \ \mathbf{B} &\triangleq \mathit{diag}(\mathbf{P})\mathbf{F}\left[\mathit{diag}(\mathbf{F}^T\mathbf{P})\right]^{-1}\mathbf{F}^T\mathit{diag}(\mathbf{P})\end{aligned} \tag{86}$$

Without further assumptions about the nature of the data, it is difficult to further simplify this equation, but it is smaller than the model utility *by definition*. Intuitively, the model utility is measuring the predictive power of the categorization (or the CA-chain model), assuming that the category classification makes no error. On the other hand, the real empirical utility is measuring the predictive power of the CA-chain model applied to the real data distribution, which is affected by the category classification error encoded in the confusion matrix. This is indeed the performance of affordance prediction in applying the CA-chain model on real data, therefore:

Claim 5 *The learning goal of categorization based affordance classification is to maximize the* empirical utility *but not* model utility.

The category recognition error is dependent on, but not limited to the following aspects:

1. *Target categorization complexity:* This measures the complexity of the "ground-truth" categorization that the agent (i.e. learner) is supposed to learn. Large number of categories and the (arbitrary) assignment of similar instances to different categories both contributes to the complexity.

2. *Agent learning capability:* The agent's learning capacity includes both the hardware components such as sensors and the software components such as the features, models and training algorithms.

3. *Amount of training data:* As previously discussed, scarce training data may be a biased sampling of real data, therefore negatively affecting the expected affordance prediction performance of the learned categorization model.

The important implication of empirical utility is that even if a categorization is indeed optimal (as measured by the model utility), it may not be the best categorization to "teach" an agent because of its incapability to learn. It is reasonable but non-trivial to conclude that the optimal categorization to provide for training should be designed based on the agent's learning capability. This also explains why the category utility should be penalized

by the number of categories as it increase the target categorization complexity for the learner. Since the number of categories affects the confusion matrix in recognition encoded in the empirical utility, it might not be necessary to use the factor $\frac{1}{N}$ (as in [27]) to prevent over-fitting.

CHAPTER VIII

CONCLUSIONS AND FUTURE DIRECTIONS

8.1 Conclusions and Contributions

In this dissertation, we have demonstrate the use of object categorization for affordance learning through our development of a computational model and further applying it to experiments in a constrained environment. More specifically, we proposed the Category-Affordance Perception model to describe the way that affordances are defined between the agent and an object, as well as the their connection to the agent's perception of the object's appearance. Mathematical models derived from this understanding is presented as in the form of graphical models with object category being a latent variable to connect the appearance and affordances. We detailed the training algorithm and presented experimental results with an indoor robot, in a number of tasks of affordance learning and prediction. Comparisons of the direct perception approach and the categorical approach are presented throughout the experiments.

The major contributions of this work are listed as follows:

1. We demonstrate that a number of tasks in robotics can be framed as affordance learning. We provide two examples of learning traversability and preferability. In traversability learning, we introduced a novel approach of robot learning affordances through autonomous exploration of the environment, such both the ambiguity of traversability definition and the need of manual data labeling are reduced. In the second task, we show that the task of "learning from examples" can be casted as preferability learning. Both tasks are successfully tested in the challenging test scenarios in the DARPA LAGR program.

2. We developed a computational model that learns and predicts affordances via object categorization. Our Category-Affordance (CA) model generally incorporates the conventional direct perception approach as well as the categorical (CA-chain and CA-full)

models that we proposed. Through theoretical discussion and experimental validation, we demonstrate the advantage of using object categorization for affordance learning compared with DP in a number of tasks, such as new affordance learning and group affordance prediction that are previously overlooked but explicitly identified in this work.

3. We draw a direct connection between the cognitive science concept of category utility originally proposed by Gluck & Corter and the application of our CA-chain model in the task of affordance learning. We show that affordances can be used as a guideline for defining object categories, i.e. object categorization. In this dissertation, we proved that the model category utility is an upper-bound of affordance prediction error under the assumption of zero category recognition error. We then introduce empirical category utility that also incorporates category recognition error, which realistically measures the actual predictive power of an object categorization for affordances prediction.

8.1.1 Application to Next Generation Personal Robots

After our discussions of using object categorization for affordance learning, we come back to the problem that we discussed earlier in the introduction chapter. The discussion in chapter 3 has justified the importance of learning rather than manually specifying affordances. Therefore, a newly purchased next generation robot can be deployed at home or in the office after a configuration stage that learns to acquire or update its knowledge of the relevant affordances. The conventional direct perception approach, although suitable for one affordance learning with large amount of data, is not generally applicable to multiple affordances learning with scarce data. Our experiments and discussions in chapters 4 to 6 suggest that object categorization is indeed useful for adaptively learning affordances. On the other hand, although categories are important to learn, the robot does not need to learn the ones that perfectly match with the linguistic categories. For example, the robot that we described is not required to know different kind of shoes or clothes, it also may ignore the difference between computer mouse, books or wallets as long as the only task that it

may do is to pick them up from the floor, in which case it only cares about whether the object is liftable.

Targeting the ultimate goal of affordance prediction, object categorization serves as an intermediate knowledge representation that is free to be organized, driven by the affordances in interest. This is demonstrated with the connection to the category utility concept discussed in chapter 7. Our discussion also points out two additional issues that are relevant in terms of specifying an object categorization for the robot to learn. First, the categorization should be more general than the scope defined by the "current" set of affordances, in the sense that it should also be good (i.e. high utility) for possible new affordances that are yet to be taught. This poses a challenging requirement that the categorization should foresee possible future affordances. While this is indeed difficult to satisfy, the common linguistic object categories may be a good choice, because they are developed in the context of our daily life. The exact granularity is then up to the robot to learn in a task driven setting. Second, the categorization should also have limited complexity: with too many categories or categories that are not easily distinguishable with the sensors available may be too difficult to learn and therefore has little empirical utility – a concept we introduced in chapter 7.

Taking into account these issues and the process of producing and deploying robots, we would recommend the following model. First, robots are manufacture-trained with a categorization with pre-specified common object categories including table, chair, fridge, etc. Most basic affordances, such as traversable and movable are then trained with this particular categorization using the category-affordance model. The initial set of object categories are defined consisting of the basic type of furniture (e.g. bed, sofa, table) and the typical objects that the robot may fetch or manipulate (e.g. soda, newspaper, kitchenware). While deploying in the user-site, the user configures the robot by first updating the category-appearance models to incorporate different appearances of the object categories as they may look differently at the new environment. Then particular affordances are taught to the robot, with training data from manual labeling or supervised robot experiments, because autonomously discovery for the non-basic affordances may be disruptive or require too much time. After initial affordance models are learned, the robot can then perform tasks to

improve its model of both the category-specific affordance classifiers and finally the object categories through re-categorization.

8.2 Future Directions

8.2.1 Data Association and Affordance Attribution

In our experiments in chapter 6, we have made the assumption that the affordance is associated with the camera image that had recently been observed. An general formation of the task is to associate the affordance label collected out of an experiment to the appearance from the sensors, which is a difficult and usually case-by-case problem. Be it direct perception or the categorical approach, the question here is how to collect training data that implies a causal (or at least dependent) relationship between the observation and the affordance? In other words, which objects in previous observations should the robot attribute to for the affordance label obtained?

To illustrate the difficulty of this problem, consider a typical robot with a camera mounted at human eye height pointing straight. When the robot's forward motion is halted by some object, the robot has to attribute this non-traversable affordance to this object. However, by the time that this affordance is observed as the halting takes place, the object is no longer in the its sight. If instead the robot associates the non-traversable affordance label to the current image observation for training, then the learned connection of the two is most likely to be arbitrary. One alternative is for the robot to reason about the previous image observation to find an object to attribute to. This would require an affordance "code book" to describe the natural of the affordance as well as a model of the 3D environment (from localization and mapping). For example, non-traversable is attributed to an object blocking the wheel or any other parts of the robot body, while supportable is attributed to the (object) surface underneath the cup being placed, but not any objects in contact with the wheel.

While our framework is applicable to affordance learning in general, significant work has to be done in defining how affordance labels can be obtained and what objects or appearance to associate with — an essential step to provide the training data. One can expect this

to be a task-by-task, robot-by-robot problem which calls for carefully designed procedures and mechanisms.

8.2.2 "Discrete" vs. "Continuous" Affordances

Our work has been focused on affordances with binary values: the object either affords or does not afford a particular action possibility. For example, an object either can or cannot support a cup of coffee, there is no intermediate or gradual change of the supportiveness of an object. This abrupt difference of the *qualitative* difference in the supportive property can be considered "categorical" [30], because it marks an innate category boundary. Mathematically speaking, the affordance value is discrete.

In contrast, consider another property which measures how easy the robot can drive on sand. It can be expected that when the sand surface is hard, the robot can easily drive on top of it; on the other extreme when the sand surface is sufficiently soft the robot's wheels sink and get stuck. The intermediate between these two extremes is possible to be quantified as a degree of how traversable the sand ground is — a value that can change gradually. This "continuous" *quantification* of affordances poses additional learning issues than the discrete or categorical case.

In terms of prediction, while predicting discrete value affordance is treated as a classification problem, continuous value affordance prediction can be formed as a regression problem. Other techniques such as discretize the affordance value can reduce the problem into the discrete case. The major difficulties, however, is in the process of data collection. Instead of obtaining a binary label from the success/failure of an experiment, the robot should now be designed to access the output of experiments *quantitatively* to measure the degree of affordances. For some affordancess, this can be done by measuring some quantity that is closely correlated with the affordance. For example, the robot can measure the actual motor power from the electrical circuits to test how much energy is used in driving, an indicator of how difficult to drive on this terrain. This capability may not be available for any continuously valued affordances, and for some affordances a manual labeling of scale is necessary in supervised learning. The data collection and association problem, as

aforementioned, is largely task specific.

In terms of model training, for discrete affordances (even if not binary), the training objective of `LL` and `CLL` remains the same as in equations 25 & 26. For learning continuous affordances, however, the regression objective is changed as minimizing the sum of squares of the residual. The object categorization learning still remains the same, but the conditional probability of $p(a_n^k|c, x_n, \theta)$ is now replaced with the regression function $f^k(x_n; c, \theta)$ with an output value between 0 to 1, where f^k denotes the regression function for a^k, also note that it is indexed by the category c. Therefore the residual of regression is defined as $\left|a_n^k - \sum_c p(c|x_n, \theta) f^k(x_n; c, \theta)\right|$, such that the second term is the average of the regression functions for all the categories weighted by the category's probability. The training objective of one affordance is to minimize the following function:

$$\texttt{SSR} = \sum_n \sum_k \left| a_n^k - \sum_c p(c|x_n, \theta) f^k(x_n; c, \theta) \right|^2 \tag{87}$$

8.2.3 Adaptation to Change in the Environment

Consider a robot with a pre-learned knowledge base with a categorization representation of the world and a set of category-to-affordance classifiers for multiple affordances (including the binomial classifiers for the CA-chain model). One would expect the robot to be able to adapt to changes in the environment, either autonomously or through manual configuration. The changes can be found in the behavior of the category-to-affordance relationships or the appearance of the categories. They can also be caused by the introduction of new affordances or categories. In section 6.6, we have demonstrate the adaptation to new affordances with the CA models, hereby we discuss adaptation to other changes as well as how the changes may be detected or notified.

Category adaptation Theoretically, learning a new category is made possible by our choice of modeling the appearance distribution (for each category) with a generative model. Unlike that with a discriminative model which requires a new decision function be learned for differentiating the new category with all existing categories, a generative model can be updated by learning the appearance description of the new category, while keeping other

categories' appearance models unchanged. In practice, however, this is complicated by two issues, the use of discriminative learning and the necessity of bookkeeping the previous training data. As we have discussed in section 6.5.3, category-appearance models learned discriminatively does not have the (desirable) property of a typical generative model as being a description of the intension of the category — in the sense that the distribution of features are learned to match the physical distribution. The appearance models in a discriminatively trained categorization are only to provide the best classification performance. Therefore learning one new category may affect all other existing category models, but not limited to the appearance distribution for the new category only. This requires that all the training data for previous categorization learning be stored.

Feature adaptation Consider a different scenario where the categories do not change but their appearances change across location and time. Note that a drastic change of both categories and their appearances could probably be treated as a new learning problem in terms of a different environment and different categorization models. We are more interested in the problems that resemble the following example: an outdoor robot (such as the LAGR one) has learned a categorization of grass, bush, tree and ground in the summer, where grass is recognized as "green stuff" and known to be traversable. However, if this robot is tested in winter, when the the grass fades and turns yellow, the performance of traversability detection will be largely affected by the fact that the appearance model no long categorizes the grass category correctly. Assuming that the categories have not changed, there are two possible scenarios of feature adaptation that the robot might face. The easier one is that although the appearance model of categories (e.g. grass being green) have changed, the discriminative features (e.g. color) for categorization remain the same. The learning can be conducted by collecting the training data on the run to learn the categorization model and the affordance classifiers. Only a few category specific classifiers need to be learned, while in the other cases categories determine the affordance values as previously learned. The difficult task is when the feature space is found no longer sufficiently discriminative — for example when both grass and bush are in the yellow color in winter. It has to be

detected that the color feature should be obsoleted (or significantly underweighted) and new features should be adopted, overweighted or discovered . This calls for the necessity of *redundant* features in the original model, i.e. there has to be at least two separate sets of features each being sufficient for categorization. This same assumption from the co-training [9] framework is not only a convenience for better utilizing unlabeled data in training, but also a backup of the model in case some features do not suit the changes in the environment.

Change detection vs. notification Before adapting to the changes in the environment, the robot has to first detect the existence and the type of the changes. By noticing that the affordance prediction performance degenerates, the robot needs to reason whether the change is caused by introducing new categories, changing in the appearance, or entering a completely new environment with different object categories. Alternatively, this can be semi-supervised learning where the existence or even the type of the changes are notified by a teacher, and the robot is required to adapt to those changes. In fact, the research of possible methods and procedures to personalize and train a manufacture-configured robot is of vital importance to build and deploy robots for any practical use.

APPENDIX A

DERIVATION OF DP FROM THE CAP MODEL

We provide a more rigorous proof of how the DP model (equation 9) is derived from the CAP model in section 4.3.1.

Proof Starting from the CAP model in equation 4 and with C marginalized out, we have:

$$\begin{aligned} & P(X=x, Y=y, A=a) \\ =\ & P(X=x, Y=y) P(A=a|Y=y) \\ \triangleq\ & P(\mathcal{R}(Y)=x, Y=y) \prod_k P(\mathcal{A}^k(Y)=a^k|Y=y) \\ \approx\ & P(\mathcal{R}(Y)=x, Y=y) \prod_k P(\mathcal{A}^k(Y)=a^k|\mathcal{R}(Y)=x) \end{aligned} \tag{88}$$

Only one term of the RHS of the equation involves y. Therefore, by integrating both sides of the equation w.r.t. y in the set of $B = \{y:\ \mathcal{A}(y)=a, \mathcal{R}(y)=x\}$ that is consistent with the observations, we obtain:

$$\begin{aligned} & P(X=x, A=a) \\ =\ & \textstyle\int_{y\in B} P(X=x, Y=y, A=a)\ dy \\ =\ & \displaystyle\int_{y\in B} P(\mathcal{R}(Y)=x, Y=y) \prod_k P(\mathcal{A}^k(Y)=a^k|Y=y)\ dy \\ \approx\ & \displaystyle\int_{y\in B} P(\mathcal{R}(Y)=x, Y=y) \prod_k P(\mathcal{A}^k(Y)=a^k|\mathcal{R}(Y)=x)\ dy \\ =\ & \displaystyle\left[\int_{y\in B} P(\mathcal{R}(Y)=x, Y=y)\ dy\right] \prod_k P(\mathcal{A}^k(Y)=a^k|\mathcal{R}(Y)=x) \\ =\ & \displaystyle P(\mathcal{R}(Y)=x) \prod_k P(\mathcal{A}^k(Y)=a^k|\mathcal{R}(Y)=x) \\ \triangleq\ & \displaystyle P(X=x) \prod_k P(A^k=a^k|X=x) \end{aligned} \tag{89}$$

which is exactly the factorization of DP. ∎

APPENDIX B

DISCUSSION ON DISCRIMINATIVE TRAINING

The following discussion about the connection between maximizing CLL and minimizing classification error is inspired by the discussion on boosting [22]. Consider the binary classification problem in which x and $y = \{0, 1\}$ denote the observation and the target label respectively and $f(x) \triangleq p(y = 1|x)$ defines the classification function. Discriminative training aims at maximizing the CLL function as follows:

$$\texttt{CLL} = \sum_n \log p(y_n|x_n) \approx \mathbf{E}\left[\log p(y|x)\right] = \mathbf{E}\left[\log(1 - |y - f(x)|)\right] \tag{90}$$

The approximation implies that the summation over the training data can be considered as a Monte Carlo sampling from the data distribution to compute the expectation of $\log p(y|x)$. On the other hand, the expected error $\mathbf{E}[\varepsilon]$ is:

$$\mathbf{E}[\varepsilon] = \mathbf{E}\left[\mathbf{I}\left(p(y|x) < 0.5\right)\right] = \mathbf{E}\left[\mathbf{I}\left(|y - f(x)| > 0.5\right)\right] \tag{91}$$

The connection between the two are shown in figure 24, with $-\log_2 p(y|x)$ being an upper-bound of the classification error $\mathbf{I}\left(p(y|x) < 0.5\right)$. Therefore maximizing CLL in discriminative training equivalently minimizes the upper-bound of classification error.

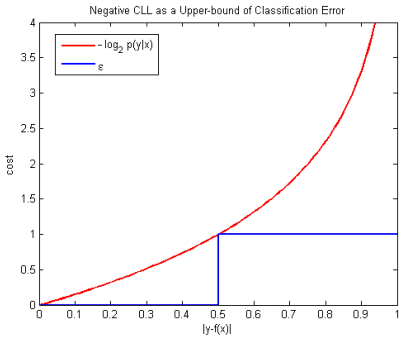


Figure 24: The negative CLL function being minimized in discriminative training is a upper-bound of expected classification error.

APPENDIX C

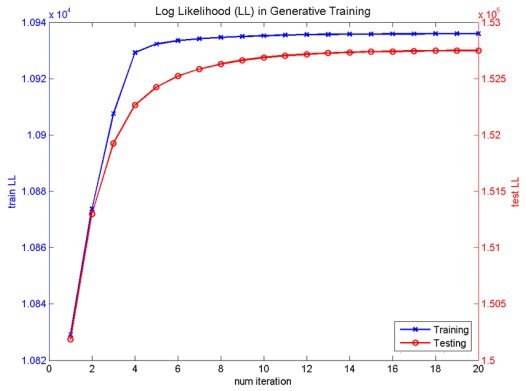
EXPERIMENTAL COMPARISON BETWEEN GENERATIVE AND DISCRIMINATIVE TRAINING

Experimental analysis of generative and discriminative training are discussed in support for previous discussion in section 6.5.1. This provides intuitive explanation of the experimental observation that models trained discriminatively achieves smaller classification error than models trained generatively.

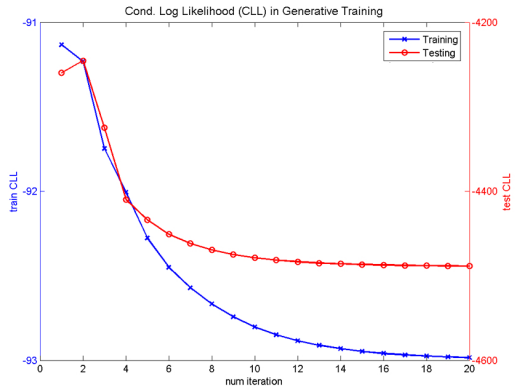
Figure 25 shows the `LL` and `CLL` functions at each EM iteration of generative training in the experiment of section 6.4. This is obtained by applying the CA-chain model at each iteration to both the training and the testing set. We see that `LL` on both the training and testing data increase at each EM iteration, but `CLL` does not always increase in training, see figure 25(b) for an example[1]. This implies that the increase in `LL` is largely due to an increase in the likelihood of appearance distributions $P(X)$ but not $P(A|X)$ that is related to the classification performance. The corresponding affordances classification error and the category recognition error both increase as in figure 25(c)(d).

On the contrary, in discriminative training the `CLL` function is maximized greedily as the training objective (figure 26(a)) which is *almost always* at the expense of a decreasing the `LL` function as our experiments suggest. Both the affordance and the category classification error decrease as with `CLL` but not `LL`. Compared to generative training, discriminative training is able to achieve an average affordance classification error of 4.5% (vs. 8.1%) and a category recognition error of 1.4% (vs. 12.9%). We also note that although the `CLL` function keeps increasing, the affordance classification error is minimized at 10^{th} iteration, justifying our post-training model selection based on affordance error.

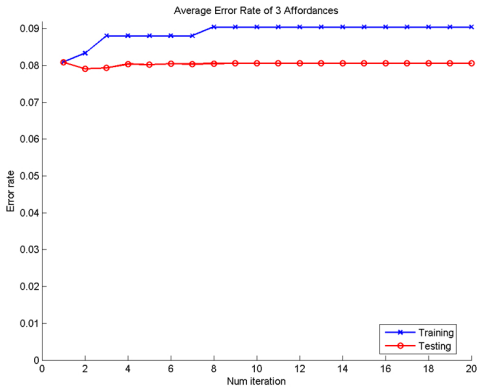
[1]It has to be noted that generative training does not always results in a decrease in the `CLL` either.



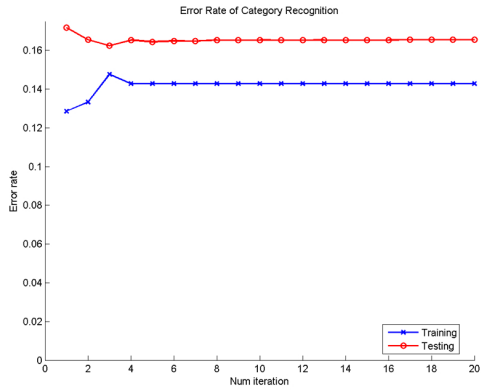
(a) `LL`



(b) `CLL`

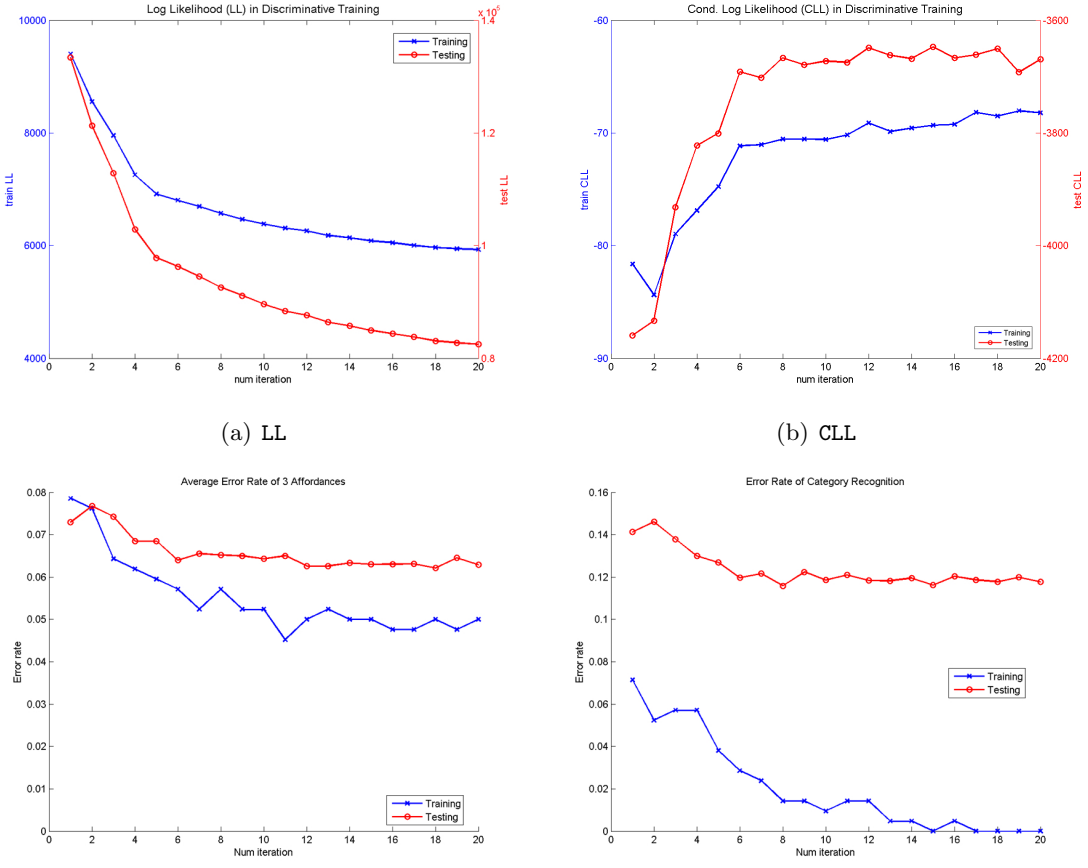


(c) Average Error



(d) Category Error

Figure 25: `LL`, `CLL`, and classification error with generative training



(a) `LL` (b) `CLL`

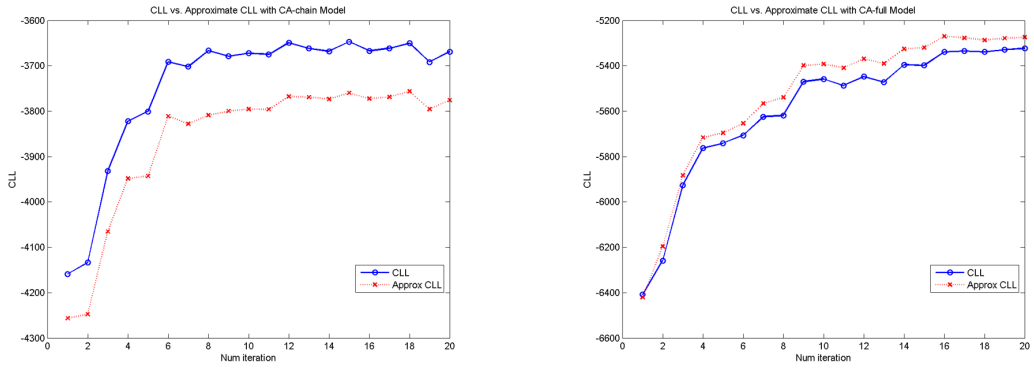
(c) Average Error (d) Category Error

Figure 26: `LL`, `CLL`, and classification error with discriminative training

APPENDIX D

DISCUSSION ON INDEPENDENT LEARNING OF CATEGORY-SPECIFIC AFFORDANCE CLASSIFIER

Here we provide evidence from our experiments in supporting for the independence approximation of category-specific affordance classifiers discussed in section 5.2. Figure 27 compares the real `CLL` function and the approximate `CLL` being maximized with independent learning of category-specific affordance classifiers[1] — in which each training sample with multiple affordance labels are treated as multiple training samples with the same appearance but only one affordance label each. For both the learned CA-chain or CA-full models, we see that the `CLL` and its approximations have the same trend in each EM iteration in training. As previously discussed in section 5.2 the approximation is valid when the mostly likely category has probability close to 1 (while other categories have probability close to 0). This is also verified in our experiment.



(a) CA-chain model (b) CA-full model

Figure 27: `CLL` vs. approximate `CLL` computed on testing set with learned CA models, suggesting that the independent learning of category-specific affordance classifiers is a good approximation of maximizing the real `CLL` function.

[1]Data obtained similarly as in section 6.5.1.

APPENDIX E

PROOF OF GLUCK UTILITY CLAIM

We prove that the Gluck utility defined in section 7.2.1 is always non-negative.

Proof We prove this by showing (a stronger result) that $\texttt{diag}(\mathbf{P}) - \mathbf{P}\mathbf{P}^T$ is semi positive definite, i.e. for any vector $\mathbf{z} \in \mathbf{R}^{\mathbf{N}}$, we show that the following is true:

$$\mathbf{z}^T \texttt{diag}(\mathbf{P})\mathbf{z} - \mathbf{z}^T\mathbf{P}\mathbf{P}^T\mathbf{z} \quad \geq \quad 0 \tag{92}$$

Next we denote the i^{th} element of $\mathbf{P}$ and $\mathbf{z}$ as p_i and z_i respectively. Note that $p_i = p(C_i)$ and hence $\sum_i p_i = 1$ by definition. The LHS of the above equation 92 can be explicitly expended as:

$$\begin{aligned}
& \mathbf{z}^T \texttt{diag}(\mathbf{P})\mathbf{z} - \mathbf{z}^T\mathbf{P}\mathbf{P}^T\mathbf{z} \\
= \quad & \sum_i p_i z_i^2 - \sum_i \sum_j p_i p_j z_i z_j \\
= \quad & \sum_i p_i z_i^2 (\sum_j p_j) - \sum_i \sum_j z_i z_j \\
= \quad & \sum_{i,j} p_i p_j z_i^2 - \sum_{i,j} p_i p_j z_i z_j \\
= \quad & \left(\sum_{i \neq j} \sum p_i p_j z_i^2 + \sum_i p_i^2 z_i^2 \right) - \left(\sum_{i \neq j} \sum p_i p_j z_i z_j + \sum_i p_i^2 z_i^2 \right) \\
= \quad & \sum_{i<j} \sum p_i p_j (z_i^2 + z_j^2) - \sum_{i<j} \sum 2 p_i p_j z_i z_j \\
= \quad & \sum_{i<j} \sum p_i p_j (z_i - z_j)^2 \\
\geq \quad & 0
\end{aligned} \tag{93}$$

Therefore $\texttt{diag}(\mathbf{P}) - \mathbf{P}\mathbf{P}^T$ is semi positive definite and hence the Gluck utility is always non-negative. ∎

APPENDIX F

PROOF OF EMPIRICAL UTILITY CLAIM

With $\mathbf{S}$, $\mathbf{P}$, $\mathbf{F}$ matrices defined previously in section 7.2.1, we prove the empirical utility in equation 86:

$$\begin{array}{rcl} \mathcal{U}_D(C;A) & \equiv & \mathtt{tr}\left(\mathbf{S}^T\mathbf{B}\mathbf{S} + (\mathbf{1}-\mathbf{S})^T\mathbf{B}(\mathbf{1}-\mathbf{S})\right) \\ where \;\; \mathbf{B} & \triangleq & \mathtt{diag}(\mathbf{P})\mathbf{F}\left[\mathtt{diag}(\mathbf{F}^T\mathbf{P})\right]^{-1}\mathbf{F}^T\mathtt{diag}(\mathbf{P}) \end{array} \tag{94}$$

Proof First, it can be shown that $\hat{\mathbf{P}}$, $\hat{\mathbf{S}}$ satisfy the following constraints:

$$\begin{cases} \mathbf{P}^T\mathbf{F} & = & \hat{\mathbf{P}}^T \\ \mathbf{P}^T\mathbf{S} & = & \hat{\mathbf{P}}^T\hat{\mathbf{S}} \end{cases} \tag{95}$$

In order to prove the claim, we note that:

$$\begin{array}{rcl} & & \mathtt{tr}\left[\mathbf{S}^T\mathbf{B}\mathbf{S} + (\mathbf{1}-\mathbf{S})^T\mathbf{B}(\mathbf{1}-\mathbf{S})\right] \\ & = & \sum_{k=1}^{K}\left[\mathbf{S_k}^T\mathbf{B}\mathbf{S_k} + (\mathbf{1}-\mathbf{S_k})^T\mathbf{B}(\mathbf{1}-\mathbf{S_k})\right] \end{array} \tag{96}$$

Therefore it suffices to show that for one affordance A^k:

$$\mathcal{U}_D(C;A^k) \;\; = \;\; \mathbf{S_k}^T\mathbf{B}\mathbf{S_k} + (\mathbf{1}-\mathbf{S_k})^T\mathbf{B}(\mathbf{1}-\mathbf{S_k}) \tag{97}$$

We assume that the predicted category $\hat{C}$ only depends on the ground truth category label C, but does not depend on the affordance in particular. Together with the fact that affordance value A^k depends on C only, we have the following factorization:

$$p(C_i, \hat{C}_j, A^k) = p(C_i)p(A^k|C_i)p(\hat{C}_j|C_i) \tag{98}$$

By Bayes law, we have:

$$\begin{array}{rcl} p(A^k=1,\hat{C}_j) & = & p(A^k=1|\hat{C}_j)p(\hat{C}_j) \\ \iff \;\; \sum_i p(C_i)p(A^k=1|C_i)p(\hat{C}_j|C_i) & = & p(A^k=1|\hat{C}_j)\sum_i p(C_i)p(\hat{C}_j|C_i) \end{array} \tag{99}$$

This can then be represented in its matrix form:

$$\mathbf{S_k}^T\mathtt{diag}(\mathbf{P})\mathbf{F_1} \;\; = \;\; p(A^k=1|\hat{C}_j)\mathbf{P}^T\mathbf{F_1} \tag{100}$$

Consider all the new categories $\hat{C}_j$ and noticing that $p(A^k = 1|\hat{C}_j)$ is the j^{th} element of vector $\hat{\mathbf{S}_\mathbf{k}}$, we have:

$$\mathbf{S_k}^T \mathtt{diag}(\mathbf{P})\mathbf{F} = \hat{\mathbf{S_k}}^T \mathtt{diag}(\mathbf{P}^T\mathbf{F}) \tag{101}$$

which implies that:

$$\hat{\mathbf{S_k}}^T = \mathbf{S_k}^T \mathtt{diag}(\mathbf{P})\mathbf{F}\left[\mathtt{diag}(\mathbf{P}^T\mathbf{F})\right]^{-1} \tag{102}$$

This together with equation 95 which states that $\mathbf{P}^T\mathbf{F} = \hat{\mathbf{P}}^T$, we have:

$$\begin{aligned}
\hat{\mathbf{S_k}}^T \mathtt{diag}(\hat{\mathbf{P}})\hat{\mathbf{S_k}} &= \mathbf{S_k}^T \mathtt{diag}(\mathbf{P})\mathbf{F}\left[\mathtt{diag}(\mathbf{P}^T\mathbf{F})\right]^{-1} \mathtt{diag}(\mathbf{P}^T\mathbf{F}) \\
&\quad \left\{\mathbf{S_k}^T \mathtt{diag}(\mathbf{P})\mathbf{F}\left[\mathtt{diag}(\mathbf{P}^T\mathbf{F})\right]^{-1}\right\}^T \\
&= \mathbf{S_k}^T \mathtt{diag}(\mathbf{P})\mathbf{F}\left[\mathtt{diag}(\mathbf{P}^T\mathbf{F})\right]^{-1} \mathbf{F}^T \mathtt{diag}(\mathbf{P})\mathbf{S_k} \\
&= \mathbf{S_k}^T \mathbf{B}\mathbf{S_k}
\end{aligned} \tag{103}$$

With the same argument for $(\mathbf{1} - \hat{\mathbf{S_k}})^T \mathtt{diag}(\hat{\mathbf{P}})(\mathbf{1} - \hat{\mathbf{S_k}})$, equation 97 is true and hence the original claim. ∎

REFERENCES

[1] *Merriam-Webster online Dictionary*. http://www.merriam-webster.com, 2008.

[2] *MobleRobots*. http://www.activrobots.com/ROBOTS/peoplebot.html, 2008.

[3] Angelova, A., *Visual Prediction of Rover Slip: Learning Algorithms and Field Experiments*. PhD thesis, California Institute of Technology, 2008.

[4] Angelova, A., Matthies, L., Helmick, D., and Perona, P., "Learning slip behavior using automatic mechanical supervision," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2007.

[5] Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I., "Matching words and pictures," *Journal of Machine Learning Research*, 2003.

[6] Bellutta, P., Manduchi, R., Matthies, L., Owens, K., and Rankin, A., "Terrain perception for demo iii," in *Intelligent Vehicle Symposium*, 2000.

[7] Berg, A., Berg, T., and Malik, J., "Shape matching and object recognition using low distortion correspondences," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[8] Blei, D. M., Ng, A. Y., and Jordan, M. I., "Latent dirichlet allocation," in *Journal of Machine Learning Research*, 2003.

[9] Blum, A. and Mitchell, T., "Combining labeled and unlabeled data with co-training," in *Workshop on Computational Learning Theory (COLT)*, 2003.

[10] Bobick, A., *Natural Object Categorization*. PhD thesis, MIT, 1987.

[11] Chemero, A. and Turvey, M. T., "Gibsonian affordances for roboticists," *Adaptive Behavior*, 2007.

[12] Corter, J. E. and Gluck, M. A., "Explaining basic categories: Feature predictability and information," in *Psychological Bulletin*, 1992.

[13] Dogar, M., Cakmak, M., Ugur, E., and Sahin, E., "From primitive behaviors to goal-directed behavior using affordances," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2007.

[14] Dogar, M., Cakmak, M., Ugur, E., and Sahin, E., "The learning and use of traversability affordance using range images on a mobile robot," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2007.

[15] Dorko, G. and Schimid, C., "Object class recognition using discriminative local features," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2004.

[16] Elfes, A., "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, 1989.

[17] Everingham, M., "The 2005 pascal visual object class challenge," in *1st PASCAL Challenges Workshop*, 2005.

[18] Fei-Fei, L., Fergus, R., and Perona, P., "A bayesian hierachical model for learning natural scene categories," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[19] Fei-Fei, L., Fergus, R., and Perona, P., "One-shot learning of object categories," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2006.

[20] Fisher, D. H., "Knowledge acquisition via incremental conceptual clustering," *Machine Learning*, vol. 2, 1987.

[21] Fitzpatrick, P., Metta, G., Natale, L., Rao, S., and Sandini, G., "Learning about objects through action - initial steps towards artificial cognition," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2003.

[22] Friedman, J., Hastie, T., and Tibshirani, R., "Additive logistic regression: a statistical view of boosting," *Annals of Statistics*, 1998.

[23] Friedman, N., Geiger, D., and Goldszmidt, M., "Bayesian network classifiers," *Machine Learning*, 1997.

[24] Fritz, G., Paletta, L., Breithaupt, R., Rome, E., and Dorffner, G., "Learning predictive features in affordance based robotic perception systems," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2006.

[25] Gibson, J. J., "The theory of affordances," *Perceiving, Acting, and Knowing*, 1977.

[26] Gibson, J. J., *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin, 1979.

[27] Gluck, M. A. and Corter, J. E., "Information, uncertainty, and the utility of categories," in *Seventh Annual Conference of the Cognitive Science Society*, June 1985.

[28] Goodman, L. A. and Kruskal, W., "Measures of association for cross classifications," *Journal of American Statistical Association*, 1954.

[29] Grauman, K. and Darrell, T., "The pyramid match kernel: Discriminative classification with sets of image features," in *Intl. Conf. on Computer Vision (ICCV)*, 2005.

[30] Harnad, S., "To cognize is to categorize: Cognition is categorization," in *Handbook on Categorization* (Lefebvre, C. and Cohen, H., eds.), Elsevier, 2005.

[31] Hofmann, T., "Probabilistic latent semantic analysis," in *Uncertainity in Artificial Intelligence*, 1999.

[32] Holub, A. and Perona, P., "A discriminative framework for modelling object classes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[33] Hoogs, A., Rittscher, J., Stein, G., and Schmiederer, J., "Video content annotation using visual analysis and large semantic knowledgebase," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[34] Huertas, A., Matthies, L., and Rankin, A., "Stereo-based tree traversability analysis for autonomous off-road navigation," in *IEEE Work-shop on Applications of Computer Vision*, 2005.

[35] Jaakkola, T. S. and Haussler, D., "Exploiting generative models in discriminative classifiers," in *Advances in Neural Information Processing Systems (NIPS)*, 1998.

[36] Jackel, L. D., Krotkov, E., Perschbacher, M., Pippine, J., and Sullivan, C., "The darpa lagr program: Goals, challenges, methodology, and phase i results," *Journal of Field Robotics*, 2007.

[37] Jensen, F., *Bayesian Networks and Decision Graphs*. Information Science and Statistics, Springer, 2002.

[38] Jing, Y., Pavlovic, V., and Rehg, J. M., "Boosted bayesian network classifiers," *Machine Learning*, 2008. accepted for publication.

[39] Jochem, T. M., Pomerleau, D. A., and Thorpe, C. E., "Vision-based neural network road and intersection detection and traversal," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 1995.

[40] Jordan, M. I., "Graphical models," in *Statistical Science (Special Issue on Bayesian Statistics)*, 2002.

[41] Jordan, M. and Jacobs, R., "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, 1994.

[42] Jordan, M. I., "Why the logistic function? a tutorial discussion on probabilities and neural networks," tech. rep., MIT Computational Cognitive Science, 1995.

[43] Kim, D., Sun, J., Oh, S. M., Rehg, J. M., and Bobick, A. F., "Traversability classification using unsupervised on-line visual learning for outdoor robot navigation," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2006.

[44] Kivinen, J. J., Sudderth, E. B., and Jordan, M. I., "Learning multiscale representations of natural scenes using dirichlet processes," in *Intl. Conf. on Computer Vision (ICCV)*, 2007.

[45] Kushal, A., Schmid, C., and Ponce, J., "Flexible object models for category-level 3d object recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[46] Lasserre, J. A., Bishop, C. M., and Minka, T. P., "Principled hybrids of generative and discriminative models," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[47] Laws, K. I., "Rapid texture identification," *Proc. of SPIE - the International Society for Optical Engineering*, 1980.

[48] Lazebnik, S., Schmid, C., and Ponce, J., "Semi-local affine parts for object recognition," in *British Machine Vision Conference (BMVC)*, 2004.

[49] Lazebnik, S., Schmid, C., and Ponce, J., "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[50] Lopes, M., Melo, F. S., and Montesano, L., "Affordance learning for manipulate small objects: Affordance-based imitation learning in robots," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2007.

[51] Lowe, D., "Object recognition from local scale-invariant features," in *Intl. Conf. on Computer Vision (ICCV)*, 1999.

[52] Marszałek, M. and Schmid, C., "Semantic hierarchies for visual object recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[53] McCullagh, P. and Nelder, J. A., *Generalized Linear Models*. Chapman & Hall, 1989.

[54] Mervis, C. B. and Rosch, E., "Categorization of natural objects," *Annual Review of Psychology*, 1981.

[55] Michels, J., Saxena, A., and Ng, A. Y., "High speed obstacle avoidance using monocular vision and reinforcement learning," in *International Conference on Machine Learning (ICML)*, 2005.

[56] Minka, T., "Discriminative models, not discriminative training," tech. rep., Microsoft Research, 2005.

[57] Mirkin, B., "Reinterpreting the category utility function," *Machine Learning*, 2001.

[58] Montesano, L., Lopes, M., Bernardino, A., and Santos-Victor, J., "Modeling object affordances using bayesian networks," in *IROS*, 2007.

[59] Montesano, L., Lopes, M., Bernardino, A., and Santos-Victor, J., "Learning object affordances: From sensory motor maps to imitation," *IEEE Transactions on Robotics*, 2008.

[60] Neisser, U., "Multiple systems: A new approach to cognitive theory," *European Journal of Cognitive Psychology*, 1994.

[61] Ng, A. Y. and Jordan, M. I., "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in Neural Information Processing Systems (NIPS)*, 2001.

[62] Nister, D. and Stewenius, H., "Scalable recognition with a vocabulary tree," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[63] Norman, D., *The Design of Everyday Things*. 1990.

[64] Opelt, A., Pinz, A., Fussenegger, M., and Auer, P., "Generic object recognition with boosting," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2006.

[65] Pagnot, R. and Grandjea, P., "Fast cross-country navigation on fair terrains," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 1995.

[66] P.Felzenschwalb and Huttenlocher, D., "Pictoral structures for object recognition," *Intl. Journal of Computer Vision*, 2005.

[67] Pinar Duygulu, Kobus Barnard, N. d. and Forsyth, D., "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *ECCV*, 2002.

[68] Pomerleau, D., "Alvinn: An autonomous land vehicle in an neural network," in *Advances in Neural Information Processing Systems (NIPS)*, 1989.

[69] Pomerleau, D., "Efficient training of artificial neural networks for autonomous navigation," *Neural Computation*, vol. 3, no. 1, 1991.

[70] Rieder, A., Southall, B., Salgian, G., Mandelbaum, R., Herman, H., Rander, P., and Stentz, T., "Stereo perception on an off-road vehicle," in *Proceedings of the Intelligent Vehicles*, 2002.

[71] Rome, E., Hertzberg, J., Dorffner, G., and Doherty, P., "Towards affordance-based robot control - abstracts collection of the dagstuhl seminar," in *Online proceedings of Dagstuhl Seminar*, 2006.

[72] Rothganger, F., Lazebnik, S., Schmid, C., and Ponce, J., "3d object modeling and recognition using local affine invariant image descriptors and multi-view spatial constraints," *Intl. Journal of Computer Vision*, 2006.

[73] Russell, B. C., Efros, A., Sivic, J., Freeman, W. T., and Zisserman., A., "Using multiple segmentations to discover objects and their extent in image collections," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[74] Russell, S. and Norvig, P., *Artificial Intelligence, A Modern Approach*. Prentice-Hall, 2002.

[75] Sahin, E., Cakmak, M., Dogar, M., Ugur, E., and G, U., "To afford or not to afford: A new formalization of affordances towards affordance-based robot control," *Adaptive Behavior*, 2007.

[76] Savarese, S. and Fei-Fei, L., "3d generic object categorization, localization and pose estimation," in *Intl. Conf. on Computer Vision (ICCV)*, 2007.

[77] Schneiderman, H., "Feature-centric evaluation for efficient cascaded object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[78] Singh, S., Simmons, R., Smith, T., Stentz, A., Verma, V., Yahja, A., and Schwehr, K., "Recent progress in local and global traversability for planetary rovers," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2000.

[79] Soderkvist, O. J. O., "Computer vision classification of leaves from swedish trees," Master's thesis, Linkoping University, 2001.

[80] Stark, L. and Bowyer, K., "Achieving generalized object recognition through reasoning about association of function to structure," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1991.

[81] Stoytchev, A., "Behavior-grounded representation of tool affordances," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2005.

[82] Sun, J., Mehta, T., Wooden, D., Power, M., Rehg, J., Balch, T., and Egerstedt, M., "Learning from examples in unstructured, outdoor environments," *Journal of Field Robotics*, 2006.

[83] Sun, J., Rehg, J. M., and Bobick, A., "Learning for ground robot navigation with autonomous data collection," Tech. Rep. GIT-GVU-05-29, Georgia Institute of Technology, Atlanta, USA, 2005.

[84] Thomas, A., Ferrari, V., Liebe, B., Tuytelaars, T., Schiele, B., and Gool, L. V., "Towards multi-view object class detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[85] Torralba, A., Murphy, K., and Freeman, W., "Sharing visual features for multiclass and multiview object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[86] Torralba, A., Fergus, R., and Freeman, W. T., "Tiny images," in *MIT-CSAIL-TR-2007-04*.

[87] Ullman, S., "Against direct perception," *Behavioral and Brain Sciences*, 1981.

[88] Ulrich, I. and Nourbakhsh, I., "Appearance-based obstacle detection with monocular color vision," in *AAAI National Conf. on Artificial Intelligence*, 2000.

[89] Ulusoy, I. and Bishop, C. M., "Generative versus discriminative methods for object recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[90] Vandapel, N., Huber, D. F., Kapuria, A., and Hebert, M., "Natural terrain classification using 3-d ladar data," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2004.

[91] Veloso, M., von Hundelshausen, F., and Rybski, P. E., "Learning visual object definitions by observing human activities," in *In Proc. IEEE-RAS Intl. Conf. on Humanoid Robots*, 2005.

[92] Viola, P. and Jones, M., "Robust real-time object detection," *International Journal of Computer Vision*, 2002.

[93] Wellington, C. and Stentz, A., "Online adaptive rough-terrain navigation in vegetation," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2004.

[94] Witten, I. H. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[95] Wooden, D. and Egerstedt, M., "Oriented visibility graphs: Low-complexity planning in real-time environments," in *IEEE Conf. on Robotics and Automation*, 2006.

[96] Wooden, D., "A guide to vision-based map building," *Robotics and Automation Magazine*, vol. 13, 2006.

[97] Wu, C. F. J., "On the convergence properties of the EM algorithm," *The Annals of Statistics*, vol. 11, 1983.

[98] Wu, J., Osuntogun, B., Choudhury, T., Philipose, M., and Rehg, J. M., "A scalable approach to activity recognition based on object use," in *Intl. Conf. on Computer Vision (ICCV)*, 2007.

[99] Wu, J., Rehg, J. M., and Mullin, M., "Learning a rare event detection cascade by direct feature selection," in *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[100] Zabih, R. and Woodfill, J., "Non-parametric local transforms for computing visual correspondence," in *European Conf. on Computer Vision (ECCV)*, 1994.